Inference in a class of optimization problems: Confidence regions and finite sample bounds on errors in coverage probabilities

Joel L. Horowitz
Sokbae Lee

Inference in a class of optimization problems: Confidence regions and finite sample bounds on errors in coverage probabilities

Joel L. Horowitz[*]

Department of Economics, Northwestern University

and

Sokbae Lee

Department of Economics, Columbia University

August 2, 2021

Abstract

This paper describes three methods for carrying out non-asymptotic inference on partially identified parameters that are solutions to a class of optimization problems. Applications in which the optimization problems arise include estimation under shape restrictions, estimation of models of discrete games, and estimation based on grouped data. The partially identified parameters are characterized by restrictions that involve the unknown population means of observed random variables in addition to the structural parameters of interest. Inference consists of finding confidence intervals for the structural parameters. Our theory provides finite-sample lower bounds on the coverage probabilities of the confidence intervals under three sets of assumptions of increasing strength. With the moderate sample sizes found in most economics applications, the bounds become tighter as the assumptions strengthen. We discuss estimation of population parameters that the bounds depend on and contrast our methods with alternative methods for obtaining confidence intervals for partially identified parameters. The results of Monte Carlo experiments and empirical examples illustrate the usefulness of our method.

Keywords: partial identification, normal approximation, sub-Gaussian distribution, finite-sample bounds

1   Introduction

This paper presents three methods for carrying out non-asymptotic inference about a function of partially identified structural parameters of an econometric model. The methods apply to models that impose shape restrictions (e.g., Freyberger and Horowitz, 2015; Horowitz and Lee, 2017), a variety of partially identified models (e.g., Manski, 2007a; Tamer, 2010) that include discrete games (e.g., Ciliberto and Tamer, 2009), and models in which a continuous function is inferred from the average values of variables in a finite number of discrete groups (e.g., Blundell, Duncan, and Meghir, 1998; Kline and Tartari, 2016). The specific inference problem consists of finding upper and lower bounds on the partially identified function $f(\psi)$ under the restrictions $g_1(\psi, \mu) \leq 0$ and $g_2(\psi, \mu) = 0$, where ψ is a vector of structural parameters, μ is a vector of unknown population means of observable random variables, f is a known, real-valued function, and g_1 and g_2 are known possibly vector-valued functions. The inequality $g_1(\psi, \mu) \leq 0$ holds component-wise.

Most existing methods for inference in our framework are based on asymptotic approximations. They provide correct inference in the limit of an infinite sample size but do not provide information about the accuracy of the asymptotic approximations in finite samples. We provide three methods for obtaining finite-sample lower bounds on the coverage probability of a confidence interval for $f(\psi)$. One method uses asymptotic approximations to obtain a confidence interval. The other two methods do not use asymptotic approximations. All the methods provide information about the accuracy of finite-sample inference.

There are several approaches to carrying out non-asymptotic inference in our framework. Sometimes a statistic with a known finite-sample distribution makes finite-sample inference possible. For example, the Clopper–Pearson (1934) confidence interval for a population probability is obtained by inverting the binomial probability distribution function. Manski (2007b) used the Clopper-Pearson interval to construct finite-sample confidence sets for counterfactual choice probabilities. Our methods apply to parameters that are not neces-

sarily probabilities. A second existing method consists of using Hoeffding's inequality to obtain a confidence interval. Syrgkanis, Tamer, and Ziani (2018) used this inequality to construct a confidence interval for a partially identified population moment. Hoeffding's inequality requires the underlying random variable to have a known bounded support. Our methods do not require the underlying random variable to have a known or bounded support. Minsker (2015) developed a confidence set for a vector of population means by using a method called "median of means." The bounds provided by this method are looser than the bounds provided by our methods. In addition, Minsker's method depends on certain user-selected tuning parameters. There are no data-based, efficient ways to choose these parameters in applications.

Our first method consists of making a normal approximation to the unknown distribution of the sample average. This method makes certain assumptions about low-order moments of the underlying random variable but does not restrict its distribution in other ways. A variety of results provide finite-sample upper bounds on the errors made by normal approximations. The Berry-Esséen inequality for the average of a scalar random variable is a well-known example of such a bound. Bentkus (2003) provides a bound on the error of a multivariate normal approximation to the distribution of the sample average of a random vector. Other normal approximations for random vectors are given by Spokoiny and Zhilova (2015); Chernozhukov, Chetverikov, and Kato (2017); and Zhilova (2020); among others. Our first method uses a bound on the error of the multivariate normal approximation that is due to Raič (2019). Raič's (2019) bound is a refined and tighter version of the bound of Bentkus (2003).

The bound of Chernozhukov, Chetverikov, and Kato (2017) may be tighter than that of Raič (2019) when the dimension of μ exceeds the sample size, but the bound of Raič (2019) is tighter when the dimension of μ is small compared to the sample size, which is the case we treat in this paper. In contrast to conventional asymptotic inference approaches, our first method provides a finite-sample lower bound on the coverage probability of a confidence

interval for the partially identified function $f(\psi)$.

The bound provided by our first method is loose in samples of the moderate sizes that occur in most economics applications, though not necessarily in very large samples. This is because it places only weak restrictions on the distribution of the underlying random variable, which may be far from normal. Our second method obtains a tighter bound in moderate size samples by assuming that the distributions of the components of the possibly vector-valued underlying random variable are sub-Gaussian. The sub-Gaussian assumption places stronger restrictions on the thickness of the tails of the relevant distributions than do the assumptions of the bound based on Raič's (2019) inequality. Our third method tightens the bound obtained with our second method by assuming that if the underlying random variable is vector-valued, then its distribution is sub-Gaussian in a vector sense. This assumption is stronger than the assumption that the components of a random vector are individually sub-Gaussian. The bounds obtained with the second and third methods are identical if the underlying random variable is a scalar.

The bounds provided by all the methods depend on unknown population parameters. This dependence is unavoidable and can be removed only in special cases. The parameters of the bounds of the second and third methods can be estimated, however, which makes it possible to estimate the bounds in applications. We describe how to do this.

Our work is broadly related to the literature on inference in partially identified models. Tamer (2010), Canay and Shaikh (2017), Ho and Rosen (2017), and Molinari (2020) provide recent surveys. Chen, Christensen, and Tamer (2018) describe a Monte Carlo method for carrying out asymptotic inference for a class of models that includes our framework. Bugni, Canay, and Shi (2017) and Kaido, Molinari, and Stoye (2019) develop asymptotic inference methods for subvectors of partially identified parameters in moment inequality models. Chernozhukov, Chetverikov, and Kato (2019) and Belloni, Bugni, and Chernozhukov (2018) construct confidence regions by inverting pointwise tests of a hypothesis about the (sub)vector of parameters that are partially identified by a large number of moment in-

equalities. The inference problem we treat is different from those in the foregoing papers in that we focus on inference about parameters that are solutions to a class of optimization problems that is different from moment inequality problems. Our methods and results do not apply to moment inequalities. Kline and Tamer (2016) describe Bayesian inference in a class of models that includes a special case of the models we treat. Two more closely related papers are Hsieh, Shi, and Shum (2017) and Shi and Shum (2015), who propose a method for asymptotic inference about estimators defined by mathematical programs. However, the class of estimation problems they treat is different from ours and overlaps with ours only under highly restrictive assumptions about both classes. Hsieh, Shi, and Shum (2017) and Shi and Shum (2015) do not provide finite-sample bounds on the errors of their asymptotic approximations.

Our work is also related to the econometrics literature on finite-sample inference. Syrgkanis, Tamer, and Ziani (2018) consider finite-sample inference in auction models. Their framework and method are very different from those in this paper. In a different context, Chernozhukov, Hansen, and Jansson (2009) and Rosen and Ura (2019) propose finite-sample inference for quantile regression models and for the maximum score estimand, respectively. Their methods and the classes of models they treat are distinct from ours.

The remainder of this paper is organized as follows. Section 2 describes the inferential problem we treat, our methods for obtaining a confidence interval for $f(\psi)$, and the three methods for obtaining a finite-sample lower bound on the coverage probability of a confidence interval. Section 2 also describes two empirical studies that illustrate how the inferential problem arises in applications. Section 3 describes computational procedures for implementing our methods. Section 4 presents an empirical application of the methods. Section 5 reports the results of a Monte Carlo investigation of the numerical performance of our methods, and Section 6 gives concluding comments. The proofs of theorems are presented in Appendix A. Appendices B–D provide additional technical information about our methods, a description of Minsker's (2015) method, and an additional empirical application.

2   The Method

Section 2.1 presents an informal description of inferential problem we address. Section 2.2 gives two examples of empirical applications in which the inferential problem arises. Section 2.3 provides a formal description of our methods for constructing confidence intervals and bounds on coverage probabilities.

2.1   The Inferential Problem

Let $\{X_i : i = 1, \dots, n\}$ be an independent random sample from the distribution of the random vector $X \in \mathbb{R}^p$ for some finite $p \geq 1$. Define $\mu = \mathbb{E}(X)$ and $\Sigma = \operatorname{cov}(X)$. We assume that both exist. Let ψ be a finite-dimensional parameter and $f(\psi)$ be a real-valued, known function. We assume throughout this section that $f(\psi)$ is only partially identified by the sampling process, though our results also hold if $f(\psi)$ is point identified. We seek a confidence interval for $f(\psi)$, which we define as a data-based interval that contains $f(\psi)$ with probability exceeding a known value. Let $g_1(\psi, \mu)$ and $g_2(\psi, \mu)$ be possibly vector valued known functions satisfying $g_1(\psi, \mu) \leq 0$ and $g_2(\psi, \mu) = 0$ component-wise. Define

$$(2.1) \qquad J_+ := \max_{\psi} f(\psi) \quad \text{and} \quad J_- := \min_{\psi} f(\psi)$$

subject to the component-wise constraints:

$$(2.2) \qquad g_1(\psi, \mu) \leq 0, \ g_2(\psi, \mu) = 0, \ \psi \in \Psi,$$

where Ψ is a compact parameter set. Appendix B extends (2.1)-(2.2) to the case in which g_1 and g_2 depend on a continuous covariate in addition to ψ and μ.

   We are interested in the identification interval $J_- \leq f(\psi) \leq J_+$. However, this interval cannot be calculated in applications because μ is unknown. Therefore, we estimate μ by the

sample average $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$, and we estimate J_+ and J_- by

(2.3) $$\hat{J}_+(\bar{X}) := \max_{\psi, m} f(\psi) \ \text{ and } \ \hat{J}_-(\bar{X}) := \min_{\psi, m} f(\psi)$$

subject to the constraints

(2.4a) $$g_1(\psi, m) \leq 0, g_2(\psi, m) = 0, \psi \in \Psi,$$

(2.4b) $$n^{1/2}(\bar{X} - m) \in \mathcal{S},$$

where $\mathcal{S}$ is a set for which $n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$ with high probability. Since μ is unknown, we replace it with the variable of optimization m in (2.3)–(2.4a) but require m to satisfy (2.4b). The resulting confidence interval for $f(\psi)$ is

(2.5) $$\hat{J}_-(\bar{X}) \leq f(\psi) \leq \hat{J}_+(\bar{X}).$$

This is also a confidence interval for the identified set containing $f(\psi)$. Section 2.3 provides three different finite-sample lower bounds on the probability that this interval contains $f(\psi)$. That is, Section 2.3 provides three finite-sample lower bounds on

(2.6) $$\mathbb{P}\left[\hat{J}_-(\bar{X}) \leq J_- \leq f(\psi) \leq J_+ \leq \hat{J}_+(\bar{X})\right].$$

The three bounds correspond to increasingly strong assumptions about the distribution of X and are increasingly tight with samples of the moderate sizes found in most economics applications, though not necessarily with very large samples.

The two leading examples of $\mathcal{S}$ in constraint (2.4b) are a box and an ellipsoid. If $\mathcal{S}$ is a box, let D be a diagonal matrix whose diagonal elements are strictly positive. For example, D might be the diagonal elements of Σ if Σ is known or the diagonal elements of a consistent estimate, $\widehat{\Sigma}$, if Σ is unknown. Choose $\kappa_b(1 - \alpha)$ so that the following holds, uniformly in

$j = 1, \ldots, p$, with probability $1 - \alpha$:

$$\left| \bar{X}_j - \mu_j \right| \leq \kappa_b(1 - \alpha)(nD_j)^{-1/2},$$

where the subscript j denotes the j'th component of a vector or the (j, j) component of a matrix. In this case, (2.4b) becomes p constraints and can be viewed as a sample analog of $|\mathbb{E}(X_j) - \mu_j| \leq 0$ with a relaxed constraint on μ_j for each $j = 1, \ldots, p$. Section 3 presents methods for choosing $\kappa_b(1 - \alpha)$.

If $\mathcal{S}$ is an ellipsoid, let Υ denote a positive definite $p \times p$ matrix, possibly Σ or $\widehat{\Sigma}$ if those matrices are non-singular, or the identity matrix. Choose $\kappa_e(1 - \alpha)$ so that

$$n(\bar{X} - \mu)' \Upsilon^{-1}(\bar{X} - \mu) \leq \kappa_e(1 - \alpha)$$

with probability $1 - \alpha$. In this case, (2.4b) is a single constraint. Section 3 presents a method for choosing $\kappa_e(1 - \alpha)$. When Σ is difficult to estimate or is singular, we may use a sphere by choosing a critical value $\kappa_s(1 - \alpha)$ such that

$$n(\bar{X} - \mu)' D^{-1}(\bar{X} - \mu) \leq \kappa_s(1 - \alpha) \quad \text{or} \quad n(\bar{X} - \mu)'(\bar{X} - \mu) \leq \kappa_s(1 - \alpha)$$

with probability $1 - \alpha$. In general, the implementation of our methods is simpler if $\mathcal{S}$ is indexed by a scalar critical value $\kappa(1 - \alpha)$.

It is straightforward to allow the objective function $f(\psi)$ to depend on μ. For the lower bound $\hat{J}_-(\bar{X})$, we introduce an auxiliary variable t that acts as an upper bound on $f(\psi, \mu)$ and solve: $\min_{\psi, m, t} t$ subject to $f(\psi, m) \leq t$ and (2.4). For the upper bound $\hat{J}_+(\bar{X})$, we introduce a lower bound s on $f(\psi, \mu)$ and solve: $\max_{\psi, m, s} s$ subject to $f(\psi, m) \geq s$ and (2.4). We focus on the original form (2.3)–(2.4) in the remainder of this paper because the form with the objective function $f(\psi, \mu)$ can be rewritten in the form (2.3)–(2.4) by redefining f, g_1 and g_2.

2.2 Examples of Empirical Applications

Example 1

Blundell, Duncan, and Meghir (1998) use grouped data to estimate labor supply effects of tax reforms in the United Kingdom. To motivate our setup, we consider a simple model with which Blundell, Duncan, and Meghir (1998) explain how to use grouped data to estimate β in the following labor supply model with no income effect:

$$(2.7) \qquad\qquad h_{it} = \alpha + \beta \log w_{it} + U_{it}.$$

In this model, h_{it} and w_{it}, respectively, are hours of work and the post-tax hourly wage rate of individual i in year t, and U_{it} is an unobserved random variable that satisfies certain conditions. The parameter β is identified by a relation of the form $\beta = \beta(h_{gt}, lw_{gt})$, where h_{gt} and lw_{gt} are the mean hours and log wages in year t of individuals in group g. There are 8 groups defined by four year-of-birth cohorts and level of education. The data span the period 1978-1992.

A nonparametric version of (2.7) is $h_{it} = \xi(w_{it}) + U_{it}$, where $\xi \in \Xi$ is an unknown continuous function and Ξ is a function space. A nonparametric analog of β is the weighted average derivative

$$\tilde{\beta} = \int \frac{\partial \xi(u)}{\partial u} w(u) du,$$

where w is a non-negative weight function. The average derivative $\tilde{\beta}$ is not identified non-parametrically by the mean values of hours and wages for finitely many groups and time periods. It can be partially identified, however, by imposing a shape restriction such as weak monotonicity on the labor supply function ξ. Assume, for example, that $\mathbb{E}[h_{it} - \xi(w_{it})|g, t] = 0$. Blundell, Duncan, and Meghir (1998) set $\mathbb{E}[h_{it} - \xi(w_{it})|g, t] = a_g + m_t$, where a_g and m_t, respectively, are group and time fixed effects. These are accommodated by our framework but we do not do this in the present discussion.

The identification interval for $\tilde{\beta}$ is $\tilde{\beta}_- \leq \tilde{\beta} \leq \tilde{\beta}_+$, where

$$(2.8) \qquad \tilde{\beta}_+ = \max_{\xi \in \Xi} \int \frac{\partial \xi(u)}{\partial u} w(u) du \ \text{ and } \ \tilde{\beta}_- = \min_{\xi \in \Xi} \int \frac{\partial \xi(u)}{\partial u} w(u) du$$

subject to

$$(2.9a) \qquad \xi(w_{gt}) - \xi(w_{g't'}) \leq 0 \ \text{ if } w_{gt} < w_{g't'},$$

$$(2.9b) \qquad h_{gt} - \xi(w_{gt}) = 0.$$

The continuous mathematical programming problem (2.8)-(2.9) can be put into the finite-dimensional framework of (2.3)-(2.4) by observing that under mild conditions on Ξ, ξ can be approximated very accurately by the truncated infinite series

$$(2.10) \qquad \xi(u) \approx \sum_{j=1}^{K} \psi_j \phi_j(u),$$

where the ψ_j's are constant parameters, the ϕ_j's are basis functions for Ξ, and K is a truncation point. In an estimation setting, K can be an increasing function of the sample size, though we do not undertake this extension here. The approximation error of (2.10) can be bounded. Here, however, we assume that K is sufficiently large to make the error negligibly small. The finite-dimensional analog of (2.8)-(2.9) is

(2.11)

$$J_+ = \max_{\psi_j:j=1,\dots,K} \sum_{j=1}^{K} \psi_j \int \frac{\partial \phi_j(u)}{\partial u} w(u) du \ \text{ and } \ J_- = \min_{\psi_j:j=1,\dots,K} \sum_{j=1}^{K} \psi_j \int \frac{\partial \phi_j(u)}{\partial u} w(u) du$$

subject to

$$\sum_{j=1}^{K} \psi_j \left[\phi_j(w_{gt}) - \phi_j(w_{g't'}) \right] \leq 0 \ \text{ if } w_{gt} < w_{g't'},$$

$$h_{gt} - \sum_{j=1}^{K} \psi_j \phi_j(w_{gt}) = 0.$$

J_+ and J_- can be estimated, thereby obtaining $\hat{J}_+$ and $\hat{J}_-$, by replacing h_{gt} and w_{gt} in (2.11)-(2.12) with within-group sample averages and adding the constraint (2.4b).

Example 2

Ho and Pakes (2014, HP hereinafter) use the theory of revealed preference to develop an estimator of hospital choices by individuals. HP use data on privately insured births in California. We consider a simplified version of the HP model.

Using the notation of HP, let $p(c, h)$ denote the price an insurer is expected to pay at hospital h for a patient with medical condition c. Let i index patients. Then c_i is the medical condition of patient i, and $p(c_i, h)$ is the price an insurer is expected to pay at hospital h for patient i. Let l_i denote patient i's location, l_h hospital's location, and $d(\cdot, \cdot)$ the distance between the two locations. For hospitals $h \neq h'$, define

$$\Delta p(c_i, h, h') := p(c_i, h) - p(c_i, h') \ \text{ and } \ \Delta d(l_i, l_h, l_{h'}) := d(l_i, l_h) - d(l_i, l_{h'}).$$

That is, $\Delta p(c_i, h, h')$ is the price difference between hospitals h and h' given patient condition c_i and $\Delta d(l_i, l_h, l_{h'})$ is the distance difference between hospitals h and h' given patient location l_i. Define

$$u_{i,i'}(\psi, h, h') := \psi \left[\Delta p(c_i, h, h') + \Delta p(c_{i'}, h', h) \right] - \left[\Delta d(l_i, l_h, l_{h'}) + \Delta d(l_{i'}, l_{h'}, l_h) \right],$$

where ψ is a scalar parameter that determines price sensitivity relative to distance. Note

that the coefficient for distance is normalized to be -1. ψ is the key parameter in HP.

Define the four-dimensional vector of instruments based on distance:

$$z_{i,i'}(h,h') := \begin{pmatrix} \max\{\Delta d(l_i, l_h, l_{h'}), 0\} \\ -\min\{\Delta d(l_i, l_h, l_{h'}), 0\} \\ \max\{\Delta d(l_{i'}, l_{h'}, l_h), 0\} \\ -\min\{\Delta d(l_{i'}, l_{h'}, l_{h'}), 0\} \end{pmatrix}.$$

Here, the instruments are based on distance measures and constructed to be positive to preserve the signs of the inequalities below in (2.13).

Let $S(h, h', s)$ be the set of patients with severity s who chose hospital h but had hospital h' in their choice set. The identifying assumption in HP is that

$$(2.13) \qquad \mathbb{E}\left[u_{i,i'}(\psi, h, h') z_{i,i'}(h, h') \middle| i \in S(h, h', s), i' \in S(h', h, s) \right] \geq 0$$

for all s, h, h' such that $h \neq h'$. We can rewrite (2.13) as

$$\psi \mu_p(h, h', s) - \mu_d(h, h', s) \geq 0 \ \text{ for all } (h, h', s) \text{ such that } h \neq h',$$

where

$$\mu_p(h, h', s) := \mathbb{E}\left[z_{i,i'}(h, h')\{\Delta p(c_i, h, h') + \Delta p(c_{i'}, h', h)\} \middle| i \in S(h, h', s), i' \in S(h', h, s) \right],$$

$$\mu_d(h, h', s) := \mathbb{E}\left[z_{i,i'}(h, h')\{\Delta d(l_i, l_h, l_{h'}) + \Delta d(l_{i'}, l_{h'}, l_h)\} \middle| i \in S(h, h', s), i' \in S(h', h, s) \right].$$

To see the connection between our general framework and HP's inequality estimator, let $f(\psi) = \psi$, $\mu = (\mu_p, \mu_d)$, and g_1 be a collection of inequalities such that

$$g_1(\psi, \mu) = \mu_d(h, h', s) - \psi \mu_p(h, h', s) \leq 0 \ \text{ for all } (h, h', s) \text{ such that } h \neq h'.$$

There is no element in g_2 (no equality constraints here). Since each element in μ can be estimated by a suitable sample mean, our general framework includes HP's estimator as a special case.

2.3 Analysis

This section presents our three methods for forming finite-sample lower bounds on

$$\mathbb{P}\left[\hat{J}_-(\bar{X}) \le J_- \le f(\psi) \le J_+ \le \hat{J}_+(\bar{X})\right].$$

The three bounds make assumptions of differing strengths about the distribution of X. The bounds are tighter in samples of moderate size with stronger assumptions. All proofs are in Appendix A. We begin with the following theorem, which applies to all the methods and forms the basis of our approach.

Theorem 2.1. *Assume that $g_1(\psi, \mu) \le 0$ and $g_2(\psi, \mu) = 0$ for some ψ. Then*

(2.14) $\qquad \mathbb{P}\left[\hat{J}_-(\bar{X}) \le J_- \le f(\psi) \le J_+ \le \hat{J}_+(\bar{X})\right] \ge \mathbb{P}\left[n^{1/2}(\bar{X} - \mu) \in \mathcal{S}\right].$

Now define

$$Z_i := X_i - \mu \quad \text{and} \quad \bar{Z} := n^{-1/2}\sum_{i=1}^{n} Z_i = n^{-1/2}\sum_{i=1}^{n}(X_i - \mu).$$

Then $\mathbb{E}(\bar{Z}) = 0$. Note that $\Sigma = \mathrm{cov}(Z_i) = \mathrm{cov}(\bar{Z}) = \mathrm{cov}(X)$. We make the following assumption throughout the remainder of the paper.

Assumption 1. *(i) $\{X_i : i = 1, \ldots, n\}$ is an independent random sample from the distribution of X. (ii) $\mathcal{S}$ is compact and convex. (iii) Ψ is compact. (iv) $f(\psi)$ is bounded on Ψ. (v) $g_1(\psi, \mu) \le 0$ and $g_2(\psi, \mu) = 0$ for some ψ.*

2.3.1 Known Σ

Suppose for the moment that Σ is known. Section 2.3.2 discusses the case in which Σ is unknown. If Σ is non-singular, let $[\Sigma^{-1/2}(X_i - \mu)]_j$ denote the j'th component of $\Sigma^{-1/2}(X_i - \mu)$.

Method 1

This method approximates the distribution of $\bar{Z}$ by a normal distribution. To do this, make the following assumption.

Assumption 2. *(i)* Σ *is non-singular, and its components are all finite.* *(ii)* *There is a constant* $\bar{\mu}_3 < \infty$ *such that* $\mathbb{E}\left(\left|[\Sigma^{-1/2}(X_i - \mu)]_j\right|^3\right) \leq \bar{\mu}_3$ *for all* $i = 1, \ldots, n$ *and* $j = 1, \ldots, p$.

Define the independent random p-vectors $W_i \sim N(0, \Sigma)$ $(i = 1, \ldots, n)$ and $\bar{W} := n^{-1/2}\sum_{i=1}^{n} W_i \sim N(0, \Sigma)$. The multivariate generalization of the Lindeberg-Lévy central limit theorem shows that $\bar{Z}$ is asymptotically distributed as $N(0, \Sigma)$, so the distribution of $\bar{Z}$ can be approximated by that of W. The following theorem bounds the error of this approximation.

Theorem 2.2. *Let Assumptions 1(i),(ii) and 2 hold. Then*

$$\left|\mathbb{P}(\bar{Z} \in \mathcal{S}) - \mathbb{P}(W \in \mathcal{S})\right| \leq \frac{(42p^{1/4} + 16)p^{3/2}\bar{\mu}_3}{n^{1/2}}.$$

Theorem 2.2 approximates the distribution of $\bar{Z}$ by a multivariate normal distribution and uses a multivariate generalization of the Berry-Esséen theorem (Raič, 2019) to bound the approximation error. Theorem 2.2 implies that for any $0 < \alpha < 1$,

$$(2.15) \qquad \mathbb{P}\left\{n(\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu) \leq \kappa_{\chi_p^2}(1 - \alpha)\right\} \geq (1 - \alpha) - \mathrm{B}(n, p, \bar{\mu}_3),$$

where $\kappa_{\chi_p^2}(1 - \alpha)$ is the $(1 - \alpha)$ quantile of the chi-square distribution with p degrees of

freedom and

$$\mathrm{B}(n, p, \overline{\mu}_3) := \frac{(42p^{1/4} + 16)p^{3/2}\overline{\mu}_3}{n^{1/2}}.$$

The term $\mathrm{B}(n, p, \overline{\mu}_3)$ in (2.15) is asymptotically negligible but can be large in samples of moderate size because it accommodates "worst case" distributions of $\bar{X}$ that may be far from normal. If n is large enough that $\mathrm{B}(n, p, \overline{\mu}_3) < \alpha$, then it follows from (2.15) that

(2.16) $$\mathbb{P}\left\{n(\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu) \le \kappa_{\chi_p^2}\left[1 - \alpha + \mathrm{B}(n, p, \overline{\mu}_3)\right]\right\} \ge 1 - \alpha.$$

It follows from Theorem 2.1 that (2.15) and (2.16) provide lower bounds on the coverage probabilities of confidence intervals for $f(\psi)$ when $\mathcal{S}$ is an ellipsoid.

Table 1 shows numerical values of the bound $(1 - \alpha) - \mathrm{B}(n, p, \overline{\mu}_3)$ and critical value $\kappa_{\chi_p^2}\left[1 - \alpha + \mathrm{B}(n, p, \overline{\mu}_3)\right]$ for different values of n and p at $\alpha = 0.05$ and $\overline{\mu}_3 = 2$. To have a bound close to $1 - \alpha$ and a finite critical value, n must be very large, especially if p is large. This is because (2.15) accommodates worst case distributions of $\bar{X}$. Methods 2 and 3, which are discussed next in this section, provide tighter bounds and smaller critical values when n is smaller, though Method 1 can provide a smaller critical value when n is very large and p is small. However, Method 1 is hard is hard to use in applications even when n is large if Σ and $\overline{\mu}_3$ are unknown, because the resulting bounds depend on population parameters that are difficult to estimate. This problem is discussed in Section 2.3.2.

Method 2

Method 2 obtains bounds that are much tighter than those of Method 1 when n is smaller than in Table 1, and Method 2 does not require Σ to be invertible. This is accomplished by assuming that the distributions of the components of X are sub-Gaussian. Specifically, make the following assumption.

Table 1: Values of the Bound and Critical Value

	$p = 1$	2	3	4	5
$n = 10^5$	0.58 (∞)	0.00 (∞)	0.00 (∞)	0.00 (∞)	0.00 (∞)
10^6	0.83 (∞)	0.58 (∞)	0.21 (∞)	0.00 (∞)	0.00 (∞)
10^7	0.91 (6.13)	0.83 (∞)	0.72 (∞)	0.57 (∞)	0.39 (∞)
10^8	0.94 (4.29)	0.91 (8.73)	0.88 (∞)	0.83 (∞)	0.77 (∞)
10^9	0.95 (3.97)	0.94 (6.53)	0.93 (9.21)	0.91 (12.88)	0.89 (∞)

Note. The table shows the values of the bound $\max[0, (1-\alpha) - \mathrm{B}(n, p, \overline{\mu}_3)]$ and critical value $\kappa_{\chi_p^2}[1 - \alpha + \mathrm{B}(n, p, \overline{\mu}_3)]$ (in parentheses) for different n and p with $\alpha = 0.05$ and $\overline{\mu}_3 = 2$.

Assumption 3. *Let Υ denote a non-singular, non-stochastic $p \times p$ matrix with finite elements. For any such Υ; all $i = 1, \ldots, n$; and all $j = 1, \ldots, p$; there are finite constants $\sigma_j^2 > 0$ such that the distribution of $\tilde{Z}_{ij} := [\Upsilon^{-1/2}(X_i - \mu)]_j$ is sub-Gaussian with variance proxy σ_j^2. That is, $\mathbb{E}[\exp(\lambda \tilde{Z}_{ij})] \leq \exp(\sigma_j^2 \lambda^2 / 2)$ for all $\lambda \in \mathbb{R}$; $i = 1, \ldots, n$; and $j = 1, \ldots, p$.*

Assumption 3 requires that the distribution of $\tilde{Z}_{ij}$ be thin-tailed. Sub-Gaussian random variables include Gaussian, Rademacher, and bounded random variables as special cases. See, e.g., Wainwright (2019). There is a tradeoff between Υ and σ_j^2. In particular, σ_j^2 may be larger if Υ is the $p \times p$ identity matrix than if $\Upsilon = \Sigma$ and Σ is non-singular.

Define $\sigma^2 := \max_{1 \leq j \leq p} \sigma_j^2$. The following theorem, combined with Theorem 2.1, provides provides a lower bound on a confidence interval for $f(\psi)$ when $\mathcal{S}$ is an ellipsoid.

Theorem 2.3. *Let Assumptions 1(i) and 3 hold. Then, for any $t > 0$,*

$$\mathbb{P}\left\{n(\bar{X} - \mu)'\Upsilon^{-1}(\bar{X} - \mu) > t\right\} \leq 2\sum_{j=1}^{p} \exp\left(-\frac{t}{2p\sigma_j^2}\right).$$

In addition, for any $0 < \alpha < 1$,

$$\mathbb{P}\left\{\frac{n}{\sigma^2}(\bar{X} - \mu)'\Upsilon^{-1}(\bar{X} - \mu) \leq \kappa_{SG,1}(1 - \alpha)\right\} \geq (1 - \alpha),$$

where $\kappa_{SG,1}(1 - \alpha) := 2p \cdot \log(2p/\alpha)$.

16

Theorem 2.3 and Method 2 make use of the sub-Gaussianity of $\tilde{Z}_{ij}$, whereas Theorem 2.2 and Method 1 allow the tails of the distribution of X to be thicker than sub-Gaussian tails. The critical values of Methods 1 and 2 are compared later in this section after the description of Method 3. Estimation of σ^2 is discussed in Section 2.3.2.

Method 3

Method 3 makes the stronger assumption that the distribution of X is sub-Gaussian in a vector sense. Specifically, Method 3 makes the following assumption.

Assumption 4. *Let* Υ *denote a non-singular* $p \times p$ *matrix with finite elements. There is a finite constant* $\sigma^2 > 0$ *such that*

$$\mathbb{E}\left[\exp(\lambda'\Upsilon^{-1/2}\bar{Z})\right] \leq \exp(\lambda'\lambda\sigma^2/2)$$

for all $\lambda \in \mathbb{R}^p$, *where* $\bar{Z} = n^{-1/2}\sum_{i=1}^{n}(X_i - \mu)$.

Assumption 4 is stronger than Assumption 3, because Assumption 4 requires the entire vector X to be sub-Gaussian. If X is multivariate normal and $\Upsilon = \Sigma$, Assumption 4 holds with $\sigma^2 = 1$. In general, however, it is difficult to find simple conditions under which Assumption 4 is satisfied without assuming that the elements of $\bar{Z}$ are independent of one another.

An application of Theorem 2.1 of Hsu, Kakade, and Zhang (2012) gives the following theorem which, combined with Theorem 2.2, provides a lower bound on the coverage probability of a confidence interval for $f(\psi)$ when $\mathcal{S}$ is an ellipsoid.

Theorem 2.4. *Let Assumptions 1(i) and 4 hold. Then, for any* $0 < \alpha < 1$,

$$\mathbb{P}\left\{\frac{n}{\sigma^2}(\bar{X} - \mu)'\Upsilon^{-1}(\bar{X} - \mu) \leq \kappa_{SG,2}(1 - \alpha)\right\} \geq 1 - \alpha,$$

where $\kappa_{SG,2}(1 - \alpha) := p + 2\sqrt{p \cdot \log(1/\alpha)} + 2\log(1/\alpha)$.

This theorem and Method 3 yield a smaller critical value and confidence set than Theorem 2.3 and Method 2 do, but they require the stronger Assumption 4. Table 2 shows the critical values of Methods 2 and 3 and chi-square critical values with p degrees of freedom. The critical values are all for $\alpha = 0.05$. None of the critical values depends on n. The chi-square critical value achieves an asymptotic coverage probability of 0.95 and yields the smallest confidence set but does not ensure a finite-sample coverage probability of at least 0.95. The sub-Gaussian critical values and resulting confidence sets are larger but ensure finite-sample coverage probabilities of at least 0.95 under the regularity conditions of Theorems 2.3 and 2.4. Assumption 3 is easier to satisfy, and its variance proxy is easier to estimate in applications, but the critical value of Method 2 increases more rapidly than the critical value of Method 3 as p gets large.

Table 2: Different Critical Values

p	$\kappa_{\chi_p^2}(1-\alpha)$	$\kappa_{SG,1}(1-\alpha)$	$\kappa_{SG,2}(1-\alpha)$
1	3.84	7.38	10.45
2	5.99	17.53	12.89
3	7.81	28.72	14.99
4	9.49	40.60	16.91
5	11.07	52.98	18.73

Note. The table shows different critical values at $\alpha = 0.05$. $\kappa_{\chi_p^2}(1-\alpha)$ is the chi-square critical value with p degrees of freedom; $\kappa_{SG,1}(1-\alpha) = 2p \cdot \log(2p/\alpha)$; and $\kappa_{SG,2}(1-\alpha) = p + 2\sqrt{p \cdot \log(1/\alpha)} + 2\log(1/\alpha)$.

The critical values of Methods 2 and 3 in Table 2 can be compared with those of Method 1 in Table 1. The critical value of Method 1 converges to the chi-square critical value as $n \to \infty$. The critical values of Methods 2 and 3 do not depend on n. Therefore, the critical value of Method 1 is smaller than those of Methods 2 and 3 when n is very large and p is small enough. However, the critical value of Method 1 is infinite with moderate values of n, whereas the critical values of Methods 2 and 3 are finite at all values of n.

18

2.3.2 Unknown Σ and sub-Gaussian variance proxy σ^2

In applications, Σ and the sub-Gaussian variance proxy σ^2 are unknown except in special cases. This section explains how to estimate these quantities and discusses the effect of estimation on the bounds presented in Section 2.3.1.

We begin with the bound of Theorem 2.2 and (2.15). Let $\widehat{\Sigma}$ be the following estimator of Σ:

$$\widehat{\Sigma} := n^{-1} \sum_{i=1}^{n} X_i X_i' - \bar{X}\bar{X}'.$$

Let Σ_{jk}^{-1} denote the (j,k) component of Σ^{-1}. For each $j, k = 1, \ldots, p$, let

$$\Sigma_{jj} := \mathbb{E}\left[(X_{ij} - \mu_j)^2\right] \quad \text{and} \quad \Gamma_{jk} := \mathbb{E}\left[\{(X_{ij} - \mu_j)(X_{ik} - \mu_k) - \Sigma_{jk}\}^2\right].$$

Assumption 5. *(i) There is a constant $C_\Sigma < \infty$ such that $\left|\Sigma_{jk}^{-1}\right| \leq C_\Sigma$ for each $j, k = 1, \ldots, p$. (ii) There is a finite constant $\kappa_0 \geq 1$ such that*

$$(2.17) \qquad\qquad\qquad \Sigma_{jj} \leq \kappa_0 \quad \text{and} \quad \Gamma_{jk} \leq \kappa_0$$

for every $j, k = 1, \ldots, p$. (iii) There is a finite constant κ_1 such that

$$(2.18\text{a}) \qquad\qquad\qquad\qquad \mathbb{E}\left[|X_{ij} - \mu_j|^r\right] \leq \kappa_1^{r-2} \frac{r!}{2} \Sigma_{jj},$$

$$(2.18\text{b}) \qquad\qquad \mathbb{E}\left[|(X_{ij} - \mu_j)(X_{ik} - \mu_k) - \Sigma_{jk}|^r\right] \leq \kappa_1^{r-2} \frac{r!}{2} \Gamma_{jk}$$

for every $r = 2, 3, \ldots$ and $j, k = 1, \ldots, p$.

Assumption 5 is stronger than Assumption 2. In particular, (2.18b) implies that the $(X_{ij} - \mu_j)(X_{ik} - \mu_k)$ is sub-exponential. Therefore, X_{ij} is sub-Gaussian because a random variable is sub-Gaussian if and only if its square is sub-exponential (Vershynin, 2018, Lemma

2.7.6). Also, the product of two sub-Gaussian variables is sub-exponential (Vershynin, 2018, Lemma 2.7.7). We use Assumption 5(iii) to apply Bernstein's inequality to the bound of Method 1 (see, e.g., Lemma 14.13 of Bühlmann and van de Geer, 2011).

Define the random vector $\widehat{W} \sim N(0, \widehat{\Sigma})$. We approximate the distribution of W by the distribution of $\widehat{W}$ with $\widehat{\Sigma}$ treated as a non-stochastic matrix. Define $\mathbf{P}(\mathcal{S}, \Sigma) := \mathbb{P}(W \in \mathcal{S})$ for $W \sim N(0, \Sigma)$ and

(2.19)
$$r_n(t) := 8\sqrt{\frac{2\kappa_0 t}{n}}, \quad w_n(t) := C_\Sigma p^3 2^{p+1} r_n(t), \quad \text{and} \quad \kappa^* := \min\left\{(4\kappa_1)^{-1}, 2\kappa_1^{-2}, (32\kappa_0)^{-1}\right\}.$$

The following lemma gives a finite-sample bound on the error of the approximation.

Lemma 2.1. *Let Assumptions 1, 2, and 5 hold and*

(2.20)
$$\log(2p) \leq t \leq \kappa^* n.$$

Then,
$$\left|\mathbf{P}(\mathcal{S}, \widehat{\Sigma}) - \mathbb{P}(W \in \mathcal{S})\right| \leq w_n(t)$$

with probability at least $1 - 2e^{-t}$.

The condition (2.20) is a mild technical condition that can be satisfied easily. The conclusion of Lemma 2.1 holds only if $\widehat{\Sigma}$ satisfies certain conditions that are stated in the proof of the lemma in Appendix A. These conditions are satisfied with probability at least $1 - 2e^{-t}$, not with certainty.

Define

(2.21)
$$\delta_n^* := \min_t \left[w_n(t) + 2e^{-t}\right] \quad \text{subject to (2.20)}.$$

Now combine Theorem 2.2 and Lemma 2.1 to obtain the following theorem.

Theorem 2.5. *Let Assumptions 1, 2, 5 and* (2.20) *hold. Then,*

$$\left| \mathbb{P}(\bar{Z} \in \mathcal{S}) - \mathbf{P}(\mathcal{S}, \widehat{\Sigma}) \right| \leq \frac{(42p^{1/4} + 16)p^{3/2}\overline{\mu}_3}{n^{1/2}} + \delta_n^*.$$

Theorem 2.5 provides a finite-sample upper bound on the error made by approximating $\mathbb{P}\left[n^{1/2}(\bar{X} - \mu) \in \mathcal{S}\right]$ by $\mathbf{P}(\mathcal{S}, \widehat{\Sigma})$. Combining Theorems 2.1 and 2.5 yields

Theorem 2.6. *Let Assumptions 1, 2, 5 and* (2.20) *hold. Then,*

$$(2.22) \quad \mathbb{P}\left[\hat{J}_-(\bar{X}) \leq J_- \leq f(\psi) \leq J_+ \leq \hat{J}_+(\bar{X})\right] \geq \mathbf{P}(\mathcal{S}, \widehat{\Sigma}) - \left\{\frac{(42p^{1/4} + 16)p^{3/2}\overline{\mu}_3}{n^{1/2}} + \delta_n^*\right\}.$$

Theorem 2.6 provides a finite-sample lower bound on $\mathbb{P}\left[\hat{J}_-(\bar{X}) \leq J_- \leq f(\psi) \leq J_+ \leq \hat{J}_+(\bar{X})\right]$ that takes account of random sampling error in $\hat{\Sigma}$. It is not difficult to choose $\mathcal{S}$ so that the right-hand side of (2.22) is $1 - \alpha$ for any $0 \leq \alpha \leq 1$ if n is large enough and p is small enough to make the term in square brackets on right-hand side of the inequality less than α. However, the presence of δ_n^* greatly decreases the right-hand side of (2.22) relative to what it is when Σ is known, thereby increasing the size of the confidence region $\mathcal{S}$ for any given value of α. In addition, the right-hand side of (2.22) depends on several population parameters that are difficult to estimate. Therefore, (2.22) is of limited use for applications. The bounds for Methods 2 and 3 with an unknown variance proxy also depend on unknown population parameters, but these are easier to estimate, as is discussed in Section 2.4. The dependence of finite-sample bounds on unknown population parameters is unavoidable except in special cases.

We now consider Method 2, which is the sub-Gaussian case of Assumption 3. In some special cases, the sub-Gaussian variance proxies, σ_j^2 are known. For example, if $\Upsilon = I_p$ in Theorem 2.3 and each element of $X - \mu$ is contained in $[-1, 1]$, then $\sigma_j^2 = 1$ for all j. Here, we derive bounds for the case in which $\sigma^2 = \max_{1 \leq j \leq p} \sigma_j^2$ is unknown and must be estimated.

For all $i = 1, \ldots, n$ and $j = 1, \ldots, p$, let

$$\tilde{X}_{ij} := [\Upsilon^{-1/2} X_i]_j,$$

and the sample variance of $\tilde{X}_{ij}$

$$\tilde{\Sigma}_{jj} := n^{-1} \sum_{i=1}^{n} \tilde{X}_{ij}^2 - \left(n^{-1} \sum_{i=1}^{n} \tilde{X}_{ij} \right)^2.$$

The following lemma establishes a finite-sample probability bound on the absolute difference between the sample and population variances of $\tilde{X}_{ij}$. Recall that $\tilde{Z}_{ij} = [\Upsilon^{-1/2}(X_i - \mu)]_j$ defined in Assumption 3.

Lemma 2.2. *Let Assumptions 1(i) and 3 hold. Then, for any $0 < t \leq (16\sigma_j^2)^2 \, n$,*

$$\mathbb{P}\left[\left| \tilde{\Sigma}_{jj} - \mathbb{E}[\tilde{Z}_{ij}^2] \right| > 3\left(\frac{t}{n}\right) + \left(1 + 2\left| n^{-1} \sum_{i=1}^{n} \tilde{X}_{ij} \right|\right)\left(\frac{t}{n}\right)^{1/2} \right] \leq e^{-t/(512\sigma_j^4)} + 2e^{-t/(2\sigma_j^2)}.$$

The next lemma establishes a link between the population variance of $\tilde{Z}_{ij}$ to the variance proxy σ_j^2.

Lemma 2.3. *Let Assumptions 1(i) and 3 hold. Then, the population variance $\mathbb{E}[\tilde{Z}_{ij}^2]$ is also the variance proxy σ_j^2.*

Now define

$$(2.23) \qquad \hat{\sigma}^2(t) := \max_{j=1,\ldots,p} \left\{ \tilde{\Sigma}_{jj} + 3\left(\frac{t}{n}\right) + \left(1 + 2\left| n^{-1} \sum_{i=1}^{n} \tilde{X}_{ij} \right|\right)\left(\frac{t}{n}\right)^{1/2} \right\},$$

which we use to estimate $\sigma^2 = \max_{1 \leq j \leq p} \sigma_j^2$. The following theorem holds for the sub-Gaussian case of Assumption 3 (Method 2).

Theorem 2.7. *Let Assumptions 1(i),(iii),(iv),(v), and 3 hold. Let the restriction $n^{1/2}(\bar{X} - $*

$\mu) \in \mathcal{S}$ *be given by*

(2.24) $$n(\bar{X} - \mu)'\Upsilon^{-1}(\bar{X} - \mu) \leq 2\hat{\sigma}^2(t) \cdot p \cdot \log(2p/\alpha),$$

where Υ is non-random and and $\hat{\sigma}^2(t)$ is given in (2.23). Then, for any $0 < t \leq (16\sigma_j^2)^2\, n$,

(2.25) $$\mathbb{P}\left[\hat{J}_-(\bar{X}) \leq J_- \leq f(\psi) \leq J_+ \leq \hat{J}_+(\bar{X})\right] \geq 1 - \alpha - \sum_{j=1}^{p}\left(e^{-t/(512\sigma_j^4)} + 2e^{-t/(2\sigma_j^2)}\right).$$

A smaller choice of t in (2.23) makes the confidence set in (2.24) tighter but results in a lower probability bound (2.25). The right-hand side of (2.25) can be close to zero and depends on the unknown parameters σ_j^2. Therefore, the bound (2.25), like the bound (2.22), is of limited use for applications. As is noted in the discussion of (2.22), dependence of finite-sample bounds on unknown population quantities is unavoidable except in special cases. Section 2.4 describes a practical approach to dealing with this problem that can be used in applications.

A result similar to Theorem 2.7 can be obtained for Method 3. We do not undertake this here, however, because the resulting bound analogous to (2.25) is loose, depends on unknown population parameters, and is of limited use in applications. Instead, in Section 2.4, we describe a way of dealing with unknown population parameters in Methods 2 and 3 that can be used in applications.

2.4 Dealing with Unknown Population Parameters in Methods 2 and 3 in Applications

Except in special cases, it is not possible to obtain finite-sample inequalities for Methods 2 and 3 that do not depend on unknown population parameters. It is possible, however, to estimate these parameters consistently and obtain asymptotic confidence intervals for them. It follows from Lemma 2.3 that $\tilde{\Sigma}_{jj}$ and $\max_{1 \leq j \leq p} \tilde{\Sigma}_{jj}$, respectively, are consistent

estimates of the variance proxies σ_j^2 and σ^2 of Method 2. Arguments like those used to prove Lemma 2.3 show that the largest eigenvalue of the covariance matrix of $\tilde{Z}_{ij}(j = 1, \ldots, p)$ is the variance proxy of Method 3. This can be estimated consistently by the largest eigenvalue of the sample covariance matrix.

Standard methods can be used to obtain asymptotic confidence intervals for the variance proxies of Method 2. These methods do not provide information about differences between true and nominal coverage probabilities in finite samples, but they provide practical indications of the magnitudes of the Method 2 bounds that can be implemented in applications.

Obtaining a useful asymptotic confidence interval for the largest eigenvalue of the covariance matrix of $\tilde{Z}_{ij}$ is difficult. A wide interval whose true coverage probability is likely to be much greater than the nominal probability can be obtained from the Frobenius norm of the difference between the estimated and true covariance matrices.

Table 4 in Section 5 compares values of $\hat{J}_+$ obtained using estimates and population values of the variance proxies of Methods 2 and 3.

3   Computational Algorithms

Recall that our general framework is to obtain the bound $[\min_{\psi,m} f(\psi), \max_{\psi,m} f(\psi)]$ subject to $g_1(\psi, m) \leq 0, g_2(\psi, m) = 0, \psi \in \Psi$, and $n^{1/2}(\bar{X} - m) \in \mathcal{S}$. In many examples, $f(\psi)$ is linear in ψ. For example, ψ is the vector of all the parameters in an econometric model and $f(\psi)$ is just one element of ψ or a linear combination of elements of ψ.

The restrictions $g_1(\psi, \mu) \leq 0$ include shape restrictions among the elements of ψ. Equality restrictions are imposed via $g_2(\psi, \mu) = 0$. The easiest case is that $g_j(\psi, \mu)$ is linear in (ψ, μ) for each $j = 1, 2$. In some of examples we consider, $g_j(\psi, \mu)$ is linear in ψ, holding μ fixed, and linear in μ, keeping ψ fixed, but not linear in (ψ, μ) jointly. This corresponds to the case of bilinear constraints. For example, $g_j(\psi, \mu)$ may depend on the product between one of elements of ψ and one of elements of μ. In practice, Ψ can always be chosen large enough

that the constraint $\psi \in \Psi$ is not binding additionally and can be ignored. For example, suppose that ψ is a probability and the constraints in $g_1(\psi, \mu) \leq 0$ and $g_2(\psi, \mu) = 0$ impose a restriction on ψ such as $[a, b]$ for $0 \leq a < b \leq 1$. Then, it is not necessary to impose $\psi \in \Psi = [0, 1]$ additionally.

Recall that the leading cases of $\mathcal{S}$ include an ellipsoid and a box. For brevity, we focus on the scenario that normal distributions are used in obtaining $\mathcal{S}$. When $\mathcal{S}$ is a box, the critical value $\kappa_b(1 - \alpha)$ can be easily simulated from the $N(0, \widehat{\Sigma})$ and the restriction $n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$ can be written as linear constraints. When $\mathcal{S}$ is an ellipsoid, the critical value $\kappa_e(1 - \alpha)$ can be obtained from the $\chi^2(d_\mu)$ distribution, where d_μ is the dimension of μ. Then, the restriction $n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$ can be written as

$$\mu' \widehat{\Sigma}^{-1} \mu - 2\mu' \widehat{\Sigma}^{-1} \bar{X} \leq n^{-1} \kappa(1 - \alpha) - \bar{X}' \widehat{\Sigma}^{-1} \bar{X}.$$

This is a convex quadratic constraint in μ.

When some of the constraints $g_1(\psi, \mu) \leq 0$ and $g_2(\psi, \mu) = 0$ are bilinear, the resulting feasible region may not be convex. To deal with the bilinear constraints, we solve optimization problems using mixed integer programming (MIP) with `Gurobi` in R. By virtue of the developments in MIP solvers and fast computing environments, MIP has become increasingly used in recent applications. For example, Bertsimas, King, and Mazumder (2016) adopted an MIO approach for obtaining ℓ_0-constrained estimators in high-dimensional regression models and Reguant (2016) used mixed integer linear programming for computing counterfactual outcomes in game theoretic models.

4   An Empirical Application

Angrist and Evans (1998) use data from the 1980 and 1990 U.S. census to estimate a model of the relation between the number of weeks per year a woman works and the number of

children she has. A simplified but nonparametric version of their model is

$$(4.1) \qquad\qquad Y = \phi(D) + U; \ \mathbb{E}[U|Z] = 0,$$

where Y is the number of weeks a woman works in a year; ϕ is an unknown function; and $D = 0, 1$, or 2 according to whether a woman has 2, 3, or 4 or more children. D is endogenous. Z is a binary instrument for D equal to 1 if the first two children are of the same sex and 0 otherwise. We obtain bounds on the partially identified parameter $\phi(0) - \phi(1)$, which measures the change in the number of weeks a woman works when the number of children she has increases from two to three. We use data consisting of 394,840 observations (Y, D, Z) from the 1980 U.S. census (Ruggles, Flood, Foster, Goeken, Pacas, Schouweiler, and Sobek, 2021).

We assume that ϕ is monotone non-increasing and focus on the parameter $f(\psi) = \phi(0) - \phi(1)$, where $\psi = [\phi(0), \phi(1), \phi(2)]'$. The population mean vector is $\mu = [\text{vec}(\Pi), \text{vec}(\mu)]$, where

$$\Pi = \begin{pmatrix} p_{DZ}(0,0) & p_{DZ}(1,0) & p_{DZ}(2,0) \\ p_{DZ}(0,1) & p_{DZ}(1,1) & p_{DZ}(2,1) \end{pmatrix}, \qquad \nu = \begin{pmatrix} \mathbb{E}[YI(Z=0)] \\ \mathbb{E}[YI(Z=1)] \end{pmatrix},$$

$p_{DZ}(d,z) := \Pr(D = d, Z = z)$, and $I(\cdot)$ is the indicator function. The dependent variable Y is contained in the interval $[0,52]$. In the analysis described below, we divide Y by 52, so it is contained in the interval $[0, 1]$, and denote the empirical analog of μ by $\bar{X}$. The inequality constraints are

$$(4.2) \qquad\qquad g_1(\psi, \mu) = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \phi(0) \\ \phi(1) \\ \phi(2) \end{pmatrix} \leq 0.$$

The equality constraints are

$$(4.3) \qquad g_2(\psi, \mu) = \begin{pmatrix} \Pi\psi - \mu \\ \sum_{d=0}^{2} \sum_{z=0}^{1} p_{DZ}(d, z) - 1 \end{pmatrix} = 0.$$

The instrumental variable constraints in (4.3) are bilinear in the sense that $\Pi\psi$ contains 6 bilinear terms.

We estimate the following 8 identification or 95% confidence intervals for $f(\psi)$.

1. (`Sample`) The identification interval given by (2.1)-(2.2) under the assumption that the sample analogs of the components of μ equal the population values. The resulting estimated identification interval is a consistent point estimate of the population identification interval but does not take account of random sampling errors in the estimate of μ.

2. (`Box`) The 95% confidence interval given by (2.3)–(2.4) with $\mathcal{S}$ a box, the D_j's equal to the diagonal elements of the sample covariance matrix $\hat{\Sigma}$, and $\kappa_b(1-\alpha) = 2.66$, which is obtained via the method given in Section 3. This estimate treats $\hat{\Sigma}$ as if it were the population covariance matrix.

3. (`Chi-Square`) The 95% confidence interval given by (2.3)–(2.4) with $\mathcal{S}$ an ellipsoid, $\Upsilon = \hat{\Sigma}$, and $\kappa_e(1-\alpha) = 14.08$, which is the 0.95 quantile of the chi-square distribution with 7 degrees of freedom (the number of distinct components of μ taking account of the constraint that $\sum_{d=0}^{2} \sum_{z=0}^{1} p_{DZ}(d, z) = 1$). This estimate, like the previous one, treats $\hat{\Sigma}$ as if it were the population covariance matrix.

4. (`Sub-Gaussian-1`) The 95% confidence interval obtained from Method 2 with $\mathcal{S}$ an ellipsoid, $\Upsilon = \hat{\Sigma}$, $\sigma^2 = 1$, and $\kappa_{SG,1}(0.95) = 78.89$. This estimate also treats $\Upsilon = \widehat{\Sigma}$ as if it were non-random. The choices of Υ and σ^2 are motivated by consistency of $\hat{\Sigma}$ for Σ.

5. (`Sub-Gaussian-2`) The 95% confidence interval obtained from Method 2 with $\mathcal{S}$ an ellipsoid, $\Upsilon = I_7$, $\sigma^2 = 1$, and $\kappa_{SG,1}(0.95) = 78.89$. This estimate does not use $\widehat{\Sigma}$. This choices of Υ and σ^2 ensure finite sample validity because each element of $X - \mu$ is contained in the interval $[-1, 1]$.

6. (`Sub-Gaussian-3`) The 95% confidence interval obtained from Method 2 with $\mathcal{S}$ an ellipsoid, $\Upsilon = I_7$, $\sigma^2 = \hat{\sigma}^2(\log n) = 0.22$ (that is, the variance proxy is estimated by (2.23) with $t = \log n$), and $\kappa_{SG,1}(0.95) = 78.89$. This estimate also does not use $\widehat{\Sigma}$ but estimates the variance proxy instead of using the bounded support of $X - \mu$.

7. (`Vector-Sub-Gaussian`) The 95% confidence interval obtained from Method 3 with $\mathcal{S}$ an ellipsoid, $\Upsilon = \hat{\Sigma}$, $\sigma^2 = 1$, and $\kappa_{SG,2}(0.95) = 22.15$. This estimate treats $\Upsilon = \widehat{\Sigma}$ as if it were not random. The choices of Υ and σ^2 are motivated by consistency of $\hat{\Sigma}$ for Σ.

8. (`Minsker`) The 95% confidence interval based on Minsker (2015). Specifically, it is obtained by replacing $n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$ with $(\hat{\mu}_* - \mu)'(\hat{\mu}_* - \mu) \leq r_{n,*}^2$, where $\hat{\mu}_*$ is a median-of-means estimator of μ based on coordinate-wise medians and $r_{n,*}^2$ is the critical value that ensures finite sample coverage. See Appendix C for details.

We do not consider a confidence interval based on (2.16) because $\mathrm{B}(n, p, \overline{\mu}_3) \approx 2.49 \overline{\mu}_3$ (here, $n = 394,840$ and $p = 7$) and as a result, it is extremely unlikely that $\mathrm{B}(n, p, \overline{\mu}_3) < 0.05$, although $\overline{\mu}_3$ is unknown. Computation was carried out on a MacBookPro laptop with an Apple M1 chip and 16 GB of memory.

The results, including computing times, are shown in Table 3. The intervals are in weeks, not weeks divided by 52. The estimates obtained by methods 1-8 above are labelled in typewriter font: `Sample`; `Box`; `Chi-Square`; `Sub-Gaussian-1`; `Sub-Gaussian-2`; `Sub-Gaussian-3`; `Vector-Sub-Gaussian`; and `Minsker`; respectively. As expected, the `Sample` interval is narrower than the `Box` and `Chi-Square` intervals, and the `Chi-Square` interval is narrower than the `Box` interval. The `Sub-Gaussian-1` and `Vector-Sub-Gaussian` intervals

are wider than the `Chi-Square` interval. The `Sub-Gaussian-1` interval is wider than the `Vector-Sub-Gaussian` interval, which is based on a stronger assumption about the distribution of the observed random variables. The foregoing methods are motivated by consistency of the estimates of the population parameters they depend on. They do not take account of random sampling error in the estimates. The `Sub-Gaussian-2` and `Minsker` intervals ensure finite-sample coverage probabilities of 0.95 but are wider than the other intervals. The `Sub-Gaussian-3` interval provides a tighter upper bound of 26.73 by estimating the variance proxy instead of relying on the bounded support of $X - \mu$. Only the `Sample` estimate yields an informative lower bound on $\phi(0) - \phi(1)$. Excluding the `Sample` bounds, which do not take account of random sampling error in the estimate of μ, the upper bounds indicate that an increase in the number of children a woman has from 2 to 3 reduces her annual employment at most by approximately 11-37 weeks, depending on the estimation method. All of the computing times are small.

Table 3: Angrist and Evans (1998) Revisited

$\mathcal{S}$	Bounds	Computing Time (sec)
`Sample`	$[1.44, 6.36]$	0.08
`Box`	$[0, 17.06]$	11
`Chi-Square`	$[0, 10.85]$	12
`Sub-Gaussian-1`	$[0, 17.19]$	15
`Sub-Gaussian-2`	$[0, 36.39]$	5
`Sub-Gaussian-3`	$[0, 26.73]$	12
`Vector-Sub-Gaussian`	$[0, 12.01]$	13
`Minsker`	$[0, 37.12]$	6

5   Monte Carlo Experiments

This section presents the results of Monte Carlo experiments that investigate the widths and coverage probabilities of nominal 95% confidence intervals for $f(\psi)$ that are obtained by using

the methods described in Section 2. For reasons explained in Section 5.1, we concentrate on the upper confidence limit $\hat{J}_+$ and do not investigate $\hat{J}_-$. The experiments are designed to mimic the empirical application of Section 4.

5.1   Design

Data were generated by simulation from model (4.1) as follows. Using the notation of Section 4, set

$$\psi = [\phi(0), \phi(1), \phi(2)] = \frac{1}{52}(23, 18, 16) \quad \text{and} \quad \Pi = \begin{pmatrix} 0.31 & 0.13 & 0.05 \\ 0.29 & 0.16 & 0.06 \end{pmatrix}.$$

The values of ϕ are similar to those used in Freyberger and Horowitz (2015), who considered larger support of D, and the values of Π are the same as the sample probabilities of Π in the empirical example of Section 4. Simulated values of (D, Z) were drawn from Π. The outcome variable Y was generated from
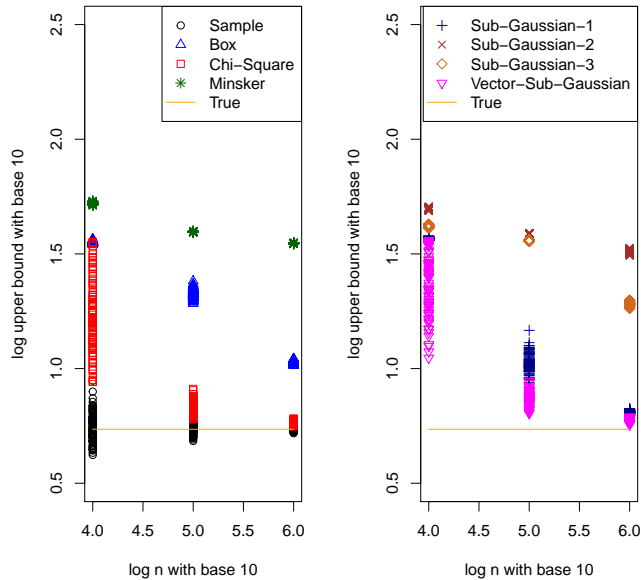
(5.1) $$Y = \mathbb{E}[\phi(D)|Z] + \frac{V}{52}; \quad V \sim N(0, 1).$$

The parameter of interest is $f(\psi) = 52[\phi(0) - \phi(1)] = 5$, which is not point-identified. We impose that each element of (ψ, μ) is non-negative and impose the monotonicity constraint (4.2). The population value of $[J_-, J_+]$ is $[0.381, 5.438]$. Thus, the data are not very informative about J_-. Consequently, we focus on $\hat{J}_+$, the estimate of J_+, in the experiments and do not investigate J_-.

We carried out experiments with $\mathcal{S}$ an ellipsoid and $\mathcal{S}$ a box. We report the averages and empirical coverage probabilities of $\hat{J}_+$ obtained by using 8 methods described in Section 4. The nominal coverage probability was 95% and there were 500 Monte Carlo replications per experiment. In experiments in which $\mathcal{S}$ is a box, $\kappa_b(1 - \alpha)$ was computed using 10^6 random

draws from $N(0, \widehat{\Sigma})$.

Figure 1: Distributions of $\hat{J}_+$

5.2 Results

Figure 1 shows only 100 (out of 500) realized values for each different estimated bound by
sample size $n \in \{10^4, 10^5, 10^6\}$. Each symbol corresponds to one Monte Carlo realization,
and the population value J_+ is shown as a horizontal line. In the figure, both x and y
axes are shown in the log scale with base 10. Table 4 summarizes the results of the Monte
Carlo experiments. The quantities in parentheses in the table are the values of $\hat{J}_+$ for the
`Sub-Gaussian-1` and `Vector-Sub-Gaussian` methods obtained with the population values
of the sub-Gaussian variance proxies, not estimates.

The `Sample` bounds converge to the true upper bound as n increases; however, they are
not suitable for inference since sampling errors are ignored. All bounds get smaller as n gets
large, but the `Chi-Square`, `Sub-Gaussian-1`, and `Vector-Sub-Gaussian` methods provide

31

Table 4: Results of Monte Carlo Experiments

	$n = 10^4$	10^5	10^6
Averages of $\hat{J}_+$			
Sample	5.60	5.43	5.45
Box	35.61	21.12	10.70
Chi-Square	17.92	6.84	5.84
Sub-Gaussian-1	36.88	10.51	6.44
Sub-Gaussian-1 with true Σ	(36.84)		
Sub-Gaussian-2	49.75	38.73	32.37
Sub-Gaussian-3	41.78	36.09	19.07
Vector-Sub-Gaussian	25.98	7.30	5.94
Vector-Sub-Gaussian with true Σ	(25.43)		
Minsker	52.62	39.47	35.21
Proportions that $\hat{J}_+ \geq J_+ = 5.438$			
Sample	0.49	0.48	0.51
All other methods	1.00	1.00	1.00

much tighter bounds than the other methods. The values of $\hat{J}_+$ for the Sub-Gaussian-1 and Vector-Sub-Gaussian methods obtained with estimated and true values of the sub-Gaussian variance proxies are nearly equal. In every case except the Sample bounds, the estimated bounds are larger than J_+, resulting in the 100% empirical coverage probabilities. Overall, the simulation results are consistent with those of the empirical application in Section 4.

6   Conclusions

This paper has described a method for carrying out inference on partially identified parameters that are solutions to a class of optimization problems. The parameters arise, for example, in applications in which grouped data are used for estimation of a model's structural parameters. Inference consists of obtaining confidence intervals for the partially identified parameters. The paper has presented three methods for obtaining finite-sample lower bounds on the coverage probabilities of the confidence intervals. The methods correspond to three sets of assumptions of increasing strength about the underlying random variable. With the

moderate sample sizes found in most economics applications, the bounds become tighter as the assumptions strengthen. The paper has also described a computational algorithm for implementing the methods. The results of Monte Carlo experiments and an empirical example illustrate the methods' usefulness. The paper has focused on the case in which a vector of first moments is the only unknown population parameter in the population version of the optimization problem. It might be useful to extend our formulation to the case in which the population optimization problem contains other population parameters. It is an open question how to obtain finite-sample lower bounds on coverage probabilities in such a case.

A   Proofs of Theorems

Proof of Theorem 2.1. Suppose $n^{1/2}(\bar{X} - \mu)$ is in $\mathcal{S}$. Any feasible solution of (2.1)–(2.2) is also a feasible solution of (2.3)–(2.4). Therefore, the feasible region of (2.3)–(2.4) contains the feasible region of (2.1)–(2.2). Consequently,

$$\hat{J}_-(\bar{X}) \leq J_- \leq J_+ \leq \hat{J}_+(\bar{X}),$$

which in turn proves (2.14). □

Proof of Theorem 2.2. Define the random p-vector $\bar{V} := \Sigma^{-1/2}\bar{Z}$. Then $\mathbb{E}(\bar{V}) = 0$ and $\mathrm{cov}(\bar{V}) = I_{p \times p}$. Define the set

(A.1) $$\mathcal{S}_\Sigma := \left\{ \xi : \xi = \Sigma^{-1/2}\zeta; \zeta \in \mathcal{S} \right\}.$$

Then $\mathcal{S}_\Sigma$ is a convex set. Define the random vectors $U_i \sim N(0, I_{p \times p})$ and $\bar{U} := n^{1/2}\sum_{i=1}^n U_i$.

It follows from Assumption 2(ii) and the generalized Minkowski inequality that

$$
\mathbb{E}\left[\left(\sum_{j=1}^{p}[\Sigma^{-1/2}(X_i - \mu)]_j^2\right)^{3/2}\right] \le p^{3/2}\overline{\mu}_3.
$$

In addition, it follows from Theorem 1.1 of Raič (2019) that

$$
\left|\mathbb{P}(\bar{V} \in \mathcal{S}_\Sigma) - \mathbb{P}(\bar{U} \in \mathcal{S}_\Sigma)\right| \le \frac{(42p^{1/4} + 16)p^{3/2}\overline{\mu}_3}{n^{1/2}},
$$

which proves the theorem. □

Proof of Theorem 2.3. Let $\|a\|$ denote the Euclidean norm for a vector a and $\tilde{Z}_i := [\Upsilon^{-1/2}(X_i - \mu)]$. Given any $r > 0$, we have that

$$
\begin{aligned}
\mathbb{P}\left(\left\|n^{-1}\sum_{i=1}^{n}\tilde{Z}_i\right\| > r\right) &=_{(1)} \mathbb{P}\left(\sum_{j=1}^{p}\left\{n^{-1}\sum_{i=1}^{n}\tilde{Z}_{ij}\right\}^2 > r^2\right) \\
&\le_{(2)} \mathbb{P}\left(\bigcup_{j=1}^{p}\left\{n^{-1}\sum_{i=1}^{n}\tilde{Z}_{ij}\right\}^2 > \frac{r^2}{p}\right) \\
&\le_{(3)} \sum_{j=1}^{p}\mathbb{P}\left(\left\{n^{-1}\sum_{i=1}^{n}\tilde{Z}_{ij}\right\}^2 > \frac{r^2}{p}\right) \\
&=_{(4)} \sum_{j=1}^{p}\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}\tilde{Z}_{ij}\right| > \frac{r}{\sqrt{p}}\right) \\
&\le_{(5)} 2\sum_{j=1}^{p}\exp\left(-\frac{nr^2}{2p\sigma_j^2}\right),
\end{aligned}
$$

where (1) is obtained by squaring both sides of the inequality, (2) comes from the fact that if $a_1 + \ldots + a_p > c$, then $a_j > c/p$ for at least one j, (3) follows because the probability of a union of events is bounded by the sum of probabilities, (4) is by taking the square root of the both sides of the inequality, and (5) is by applying Hoeffding bound for the sum of independent sub-Gaussian random variables (see, e.g., Proposition 2.5 of Wainwright, 2019).

In particular, the result above implies that if we take $\sigma^2 = \max_{1 \leq j \leq p} \sigma_j^2$,

$$\mathbb{P}\left\{\frac{n}{\sigma^2}(\bar{X} - \mu)'\Upsilon^{-1}(\bar{X} - \mu) > \kappa_{SG,1}(1 - \alpha)\right\} \leq \alpha,$$

which proves the theorem. $\square$

Proof of Theorem 2.4. It follows from Theorem 2.1 of Hsu, Kakade, and Zhang (2012) (by taking A in their theorem to be the p-dimensional identity matrix) that

$$\mathbb{P}\left\{\frac{n}{\sigma^2}(\bar{X} - \mu)'\Upsilon^{-1}(\bar{X} - \mu) > p + 2\sqrt{pt} + 2t\right\} \leq \exp(-t)$$

for all $t > 0$. This yields the theorem immediately. $\square$

Proof of Lemma 2.1. Define

$$\Delta_n := \sup_{\mathcal{S}} \left|\mathbf{P}(\mathcal{S}, \widehat{\Sigma}) - \mathbb{P}(W \in \mathcal{S})\right|.$$

Now

$$\mathbf{P}(\mathcal{S}, \widehat{\Sigma}) - \mathbb{P}(W \in \mathcal{S}) = \mathbb{P}(\Sigma^{-1/2}\widehat{W} \in \mathcal{S}_\Sigma) - \mathbb{P}(\xi \in \mathcal{S}_\Sigma),$$

where $\xi \sim N(0, I_{p \times p})$ and $\mathcal{S}_\Sigma$ is defined in (A.1). Therefore,

$$\begin{aligned}
\Delta_n &= \sup_{\mathcal{S}} \left|\mathbb{P}(\Sigma^{-1/2}\widehat{W} \in \mathcal{S}_\Sigma) - \mathbb{P}(\xi \in \mathcal{S}_\Sigma)\right| \\
&\leq \mathrm{TV}\left[N(0, I_{p \times p}), N(0, \Sigma^{-1}\widehat{\Sigma})\right],
\end{aligned}$$

where $\mathrm{TV}(P_1, P_2)$ is the total variation distance between distributions P_1 and P_2. By Example 2.3 of Dasgupta (2008),

$$\mathrm{TV}\left[N(0, I_{p \times p}), N(0, \Sigma^{-1}\widehat{\Sigma})\right] \leq p 2^{p+1} \left\|\Sigma^{-1}\widehat{\Sigma} - I_{p \times p}\right\|_{\mathrm{F}}$$

where for any matrix A,

$$\|A\|_{\mathrm{F}}^2 := \sum_{j=1}^{p} \sum_{j=1}^{p} a_{jk}^2.$$

Define $\omega := \widehat{\Sigma} - \Sigma$. Then,

$$\Sigma^{-1}\widehat{\Sigma} - I_{p\times p} = \Sigma^{-1}(\widehat{\Sigma} - \Sigma) = \Sigma^{-1}\omega,$$

$$\left|(\Sigma^{-1}\omega)_{jk}\right| \le \sum_{\ell=1}^{p} \left|\Sigma_{j\ell}^{-1}\omega_{\ell k}\right| \le C_{\Sigma} \sum_{\ell=1}^{p} |\omega_{\ell k}|$$

and

$$\left\|\Sigma^{-1}\widehat{\Sigma} - I_{p\times p}\right\|_{\mathrm{F}} \le C_{\Sigma}p^{1/2}\left[\sum_{k=1}^{p}\left(\sum_{\ell=1}^{p}|\omega_{\ell k}|\right)^2\right]^{1/2}.$$

To obtain the conclusion of the theorem, it remains to show that $|\omega_{jk}| \le r_n(t)$ with probability at least $1 - 2e^{-t}$. We prove this claim below.

Write

$$\omega = n^{-1}\sum_{i=1}^{n}[(X_i - \mu)(X_i - \mu)' - \Sigma] - (\bar{X} - \mu)(\bar{X} - \mu)'.$$

Then, we have that

$$\mathbb{P}\left[\max_{j,k}|\omega_{jk}| \ge r_n(t)\right]$$

$$\le \mathbb{P}\left[\max_{1\le j,k\le p}\left|n^{-1}\sum_{i=1}^{n}[(X_{ij} - \mu)(X_{ik} - \mu)' - \Sigma_{jk}]\right| \ge \frac{r_n(t)}{2}\right] + \mathbb{P}\left[\left\{\max_{1\le j\le p}|\bar{X}_j - \mu_j|\right\}^2 \ge \frac{r_n(t)}{2}\right].$$

Recall that

$$\Sigma_{jj} = \mathbb{E}\left[|X_{ij} - \mu_j|^2\right] \quad \text{and} \quad \Gamma_{jk} = \mathbb{E}\left[|(X_{ij} - \mu_j)(X_{ik} - \mu_k) - \Sigma_{jk}|^2\right].$$

Define

$$\lambda(\kappa_1, n, p) := \sqrt{\frac{2\log(2p)}{n}} + \frac{\kappa_1 \log(2p)}{n}.$$

By Bernstein's inequality for the maximum of p averages (see, e.g., Lemma 14.13 of Bühlmann and van de Geer (2011)),

$$\mathbb{P}\left[\max_{1 \le j \le p} \left|\Sigma_{jj}^{-1/2}(\bar{X}_j - \mu_j)\right| \ge \frac{\kappa_1 t}{\Sigma_{jj}^{1/2} n} + \sqrt{\frac{2t}{n}} + \lambda\left(\Sigma_{jj}^{-1/2}\kappa_1, n, p\right)\right] \le \exp\left(-t\right)$$

and

$$\mathbb{P}\left[\max_{1 \le j,k \le p} \left|\Gamma_{jk}^{-1/2} n^{-1} \sum_{i=1}^{n} [(X_{ij} - \mu)(X_{ik} - \mu)' - \Sigma_{jk}]\right| \ge \frac{\kappa_1 t}{\Gamma_{jk}^{1/2} n} + \sqrt{\frac{2t}{n}} + \lambda(\Gamma_{jk}^{-1/2}\kappa_1, n, p^2)\right] \le \exp\left(-t\right).$$

Suppose that for each j,

(A.2) $$\max\{t, \log(2p)\} \le \min\left\{\frac{n}{4\kappa_1}, \frac{n}{32\Sigma_{jj}}\right\}.$$

Under (A.2),

$$\mathbb{P}\left[\left\{\max_{1 \le j \le p} \left|\bar{X}_j - \mu_j\right|\right\}^2 \ge \frac{\kappa_1 t}{n} + \Sigma_{jj}^{1/2}\sqrt{\frac{2t}{n}} + \Sigma_{jj}^{1/2}\lambda\left(\Sigma_{jj}^{-1/2}\kappa_1, n, p\right)\right]$$

$$\le \mathbb{P}\left[\left\{\max_{1 \le j \le p} \left|\bar{X}_j - \mu_j\right|\right\}^2 \ge \left\{\frac{\kappa_1 t}{n} + \Sigma_{jj}^{1/2}\sqrt{\frac{2t}{n}} + \Sigma_{jj}^{1/2}\lambda\left(\Sigma_{jj}^{-1/2}\kappa_1, n, p\right)\right\}^2\right]$$

$$\le \exp\left(-t\right).$$

Therefore, if we can choose $r_n(t)$ by

(A.3) $$\frac{r_n(t)}{2} := \max_{j,k}\left\{\Sigma_{jj}^{1/2}, \Gamma_{jk}^{1/2}\right\}\left\{\sqrt{\frac{2t}{n}} + \sqrt{\frac{2\log(2p)}{n}}\right\} + \frac{\kappa_1 t}{n} + \frac{\kappa_1 \log(2p)}{n}$$

for $t > 0$, we have that

$$\mathbb{P}\left[\max_{j,k} |\omega_{jk}| \le r_n(t)\right] \ge 1 - 2e^{-t}.$$

To simplify the upper bound $r_n(t)$, we now strengthen (A.2) to (2.20) stated in the theorem.
Then, we can take

$$r_n(t) := 8\sqrt{\frac{2\kappa_0 t}{n}},$$

which proves the claim. $\qquad\square$

Proof of Theorem 2.5. Write

$$\left|\mathbb{P}(\bar{Z} \in \mathcal{S}) - \mathbf{P}(\mathcal{S}, \widehat{\Sigma})\right| = \left|\left[\mathbb{P}(\bar{Z} \in \mathcal{S}) - \mathbb{P}(W \in \mathcal{S})\right] - \left[\mathbf{P}(\mathcal{S}, \widehat{\Sigma}) - \mathbb{P}(W \in \mathcal{S})\right]\right|$$

$$\le \left|\mathbb{P}(\bar{Z} \in \mathcal{S}) - \mathbb{P}(W \in \mathcal{S})\right| + \left|\mathbf{P}(\mathcal{S}, \widehat{\Sigma}) - \mathbb{P}(W \in \mathcal{S})\right|.$$

Thus, the theorem follows immediately by combining Theorem 2.2 and Lemma 2.1. $\qquad\square$

Proof of Theorem 2.6. Combining Theorems 2.1 and 2.5 yields Theorem 2.6. $\qquad\square$

Proof of Lemma 2.2. By Lemma 1.12 of Rigollet and Hütter (2017), $(\tilde{X}_{ij}^2 - \mathbb{E}[\tilde{X}_{ij}^2])$ is sub-exponential with parameter $16\sigma_j^2$. This in turn implies that by Theorem 1.13 of Rigollet and Hütter (2017), which is a version of Bernsteins inequality, we have, for any $s > 0$,

$$\mathbb{P}\left[\left|n^{-1}\sum_{i=1}^{n} \tilde{X}_{ij}^2 - \mathbb{E}[\tilde{X}_{ij}^2]\right| > s\right] \le \exp\left[-\frac{n}{2}\min\left\{\frac{s^2}{(16\sigma_j^2)^2}, \frac{s}{16\sigma_j^2}\right\}\right].$$

Take $s = (t/n)^{1/2}$. Then, for any $0 < t \le (16\sigma_j^2)^2 n$,

(A.4) $$\mathbb{P}\left[\left|n^{-1}\sum_{i=1}^{n} \tilde{X}_{ij}^2 - \mathbb{E}[\tilde{X}_{ij}^2]\right| > \left(\frac{t}{n}\right)^{1/2}\right] \le e^{-t/(512\sigma_j^4)}.$$

As before, by Hoeffding bound for the sum of independent sub-Gaussian random variables

(see, e.g., Proposition 2.5 of Wainwright, 2019), we have that

$$\mathbb{P}\left[\left|n^{-1}\sum_{i=1}^{n}\tilde{X}_{ij} - \mathbb{E}[\tilde{X}_{ij}]\right| \geq \left(\frac{t}{n}\right)^{1/2}\right] \leq 2e^{-t/(2\sigma_j^2)}.$$

Furthermore, using the fact that

$$\left|\left(n^{-1}\sum_{i=1}^{n}\tilde{X}_{ij}\right)^2 - (\mathbb{E}[\tilde{X}_{ij}])^2\right| \leq \left(n^{-1}\sum_{i=1}^{n}\tilde{X}_{ij} - \mathbb{E}[\tilde{X}_{ij}]\right)^2 + 2\left|\mathbb{E}[\tilde{X}_{ij}]\right|\left|n^{-1}\sum_{i=1}^{n}\tilde{X}_{ij} - \mathbb{E}[\tilde{X}_{ij}]\right|$$

$$\leq 3\left(n^{-1}\sum_{i=1}^{n}\tilde{X}_{ij} - \mathbb{E}[\tilde{X}_{ij}]\right)^2 + 2\left|n^{-1}\sum_{i=1}^{n}\tilde{X}_{ij}\right|\left|n^{-1}\sum_{i=1}^{n}\tilde{X}_{ij} - \mathbb{E}[\tilde{X}_{ij}]\right|,$$

we have that

(A.5)
$$\mathbb{P}\left[\left|\left(n^{-1}\sum_{i=1}^{n}\tilde{X}_{ij}\right)^2 - (\mathbb{E}[\tilde{X}_{ij}])^2\right| \geq 3\left(\frac{t}{n}\right) + 2\left|n^{-1}\sum_{i=1}^{n}\tilde{X}_{ij}\right|\left(\frac{t}{n}\right)^{1/2}\right]$$
$$\leq \mathbb{P}\left[\left|n^{-1}\sum_{i=1}^{n}\tilde{X}_{ij} - \mathbb{E}[\tilde{X}_{ij}]\right| \geq \left(\frac{t}{n}\right)^{1/2}\right] \leq 2e^{-t/(2\sigma_j^2)}.$$

The desired result follows by combining (A.5) with (A.4). □

Proof of Lemma 2.3. By Assumption 3, $\tilde{Z}_{ij} = [\Upsilon^{-1/2}(X_i - \mu)]_j$ is a mean-zero scalar sub-Gaussian random variable with variance proxy σ^2, which is denoted by $\tilde{Z}_{ij} \sim \mathrm{subG}(\sigma^2)$. Let $\bar{\tilde{Z}}_j := n^{-1/2}\sum_{i=1}^{n}\tilde{Z}_{ij}$. Then $\bar{\tilde{Z}}_j \to_d N(0, \mathbb{E}[\tilde{Z}_{ij}^2])$ and $\bar{\tilde{Z}}_j \sim \mathrm{subG}(\sigma^2)$ for all n. Let $\varphi_{\tilde{Z}}, \varphi_n$, and φ, respectively, denote the moment-generating functions of $\tilde{Z}_{ij}, \bar{\tilde{Z}}_j$, and $N(0, \mathbb{E}[\tilde{Z}_{ij}^2])$. Then

$$\varphi(t) = \exp\left(\mathbb{E}[\tilde{Z}_{ij}^2]t^2/2\right),$$

(A.6)
$$\varphi_n(t) \leq \exp(\sigma^2 t^2/2),$$

and

$$\varphi_n(t) = \mathbb{E}\exp(\bar{\tilde{Z}}_j t) = \mathbb{E}\exp\left(\sum_{i=1}^{n} \frac{t\tilde{Z}_{ij}}{n^{1/2}}\right) = \prod_{i=1}^{n} \mathbb{E}\exp\left(\frac{t\tilde{Z}_{ij}}{n^{1/2}}\right) = \left[\varphi_{\tilde{Z}}\left(\frac{t}{n^{1/2}}\right)\right]^n.$$

Arguments identical to those used to prove the Lindeberg-Lévy central limit theorem show that

(A.7)
$$\lim_{n\to\infty} \varphi_n(t) = \exp\left(\mathbb{E}[\tilde{Z}_{ij}^2]t^2/2\right).$$

Let $\breve{\sigma}^2$ be the smallest σ^2 for which (A.6) is satisfied. If $\breve{\sigma}^2 > \mathbb{E}[\tilde{Z}_{ij}^2]$, then (A.7) is impossible. Therefore, $\breve{\sigma}^2 \leq \mathbb{E}[\tilde{Z}_{ij}^2]$. If any random variable $Z \sim \mathrm{subG}(\breve{\sigma}^2)$ and $\mathbb{E}[Z^2] > \breve{\sigma}^2$, then $Z \sim \mathrm{subG}(\mathbb{E}[Z^2])$. Therefore, $\tilde{Z}_{ij} \sim \mathrm{subG}(\mathbb{E}[Z^2])$. $\qquad\square$

Proof of Theorem 2.7. Note that

$$\mathbb{P}\left\{n(\bar{X} - \mu)'\Upsilon^{-1}(\bar{X} - \mu) > \hat{\sigma}^2(t)\kappa_{SG,1}(1 - \alpha)\right\}$$

$$= \mathbb{P}\left\{n(\bar{X} - \mu)'\Upsilon^{-1}(\bar{X} - \mu) > \hat{\sigma}^2(t)\kappa_{SG,1}(1 - \alpha) \quad \text{and} \quad \hat{\sigma}^2(t) < \sigma^2\right\}$$

$$\quad + \mathbb{P}\left\{n(\bar{X} - \mu)'\Upsilon^{-1}(\bar{X} - \mu) > \hat{\sigma}^2(t)\kappa_{SG,1}(1 - \alpha) \quad \text{and} \quad \hat{\sigma}^2(t) \geq \sigma^2\right\}$$

$$\leq \mathbb{P}\left\{\hat{\sigma}^2(t) < \sigma^2\right\} + \mathbb{P}\left\{n(\bar{X} - \mu)'\Upsilon^{-1}(\bar{X} - \mu) > \sigma^2\kappa_{SG,1}(1 - \alpha)\right\}.$$

Then, the theorem follows from Lemma 2.2 and Theorem 2.1. $\qquad\square$

B Continuous Covariates

In this section, we consider the case in which g_1 and g_2 depend on a continuous covariate ν in addition to (ψ, μ). This situation occurs, for example, in applications where some observed variables are group averages and others are continuously distributed characteristics of individuals. If ν is discrete, the results of Section 2.3 apply after replacing problem (2.3)-

(2.4) with (B.3)-(B.4) below. When there is a continuous covariate, ν, (2.3)-(2.4) become

(B.1) $$\hat{J}_+(\bar{X}) := \max_{\psi, m} f(\psi) \quad \text{and} \quad \hat{J}_-(\bar{X}) := \min_{\psi, m} f(\psi)$$

subject to

(B.2a) $$g_1(\psi, m, \nu) \leq 0 \text{ for every } \nu,$$

(B.2b) $$g_2(\psi, m, \nu) = 0 \text{ for every } \nu,$$

(B.2c) $$\psi \in \Psi,$$

(B.2d) $$n^{1/2}(\bar{X} - m) \in \mathcal{S}.$$

Thus, there is a continuum of constraints. We form a discrete approximation to (B.2a)-(B.2b) by restricting ν to a discrete grid of points. Let L denote the number of grid points. We give conditions under which the optimal values of the objective functions of the discretized version of (B.1)-(B.2) converge to $\hat{J}_+(\bar{X})$ and $\hat{J}_-(\bar{X})$ as $L \to \infty$. To minimize the notational complexity of the following discussion we assume that ν is a scalar. The generalization to a vector is straightforward. We also assume that ν is contained in a compact set which, without further loss of generality, we take to be $[0, 1]$.

To obtain the grid approximation, let $0 = v_0 < v_1 < v_2 < \ldots < v_L = 1$ be a grid of equally space points in $[0, 1]$. The distance between grid points is $1/(L - 1)$. Approximate problem (B.1)-(B.2) by

(B.3) $$\tilde{J}_+(\bar{X}) := \max_{\psi, m} f(\psi) \quad \text{and} \quad \tilde{J}_-(\bar{X}) := \min_{\psi, m} f(\psi)$$

subject to the constraints:

(B.4a)
$$g_1(\psi, m, \nu_\ell) \leq 0; \ \ell = 1, \ldots, L,$$

(B.4b)
$$g_2(\psi, m, \nu_\ell) = 0; \ \ell = 1, \ldots, L,$$

(B.4c)
$$\psi \in \Psi,$$

and

(B.4d)
$$n^{1/2}(\bar{X} - m) \in \mathcal{S}.$$

We then have

Theorem B.1. *Assume that f is continuous, $\nu \in [0, 1]$, and m in (B.4) is contained in a compact set $\mathcal{M}$. Moreover,*

$$|g_j(\psi, m; x) - g_j(\psi, m; v_\ell)| \leq C\,|x - v_\ell|$$

$$|g_j(\psi, m; x) - g_j(\psi, m; v_{\ell+1})| \leq C\,|x - v_{\ell+1}|$$

for $j = 1$ or 2, some $C < \infty$, and all $\psi \in \Psi$, all $m \in \mathcal{M}$, all $x \in [v_\ell, v_{\ell+1}] \in [0, 1]$. Then

(B.5)
$$\lim_{L \to \infty} \tilde{J}_+ = \hat{J}_+ \ \ and \ \ \lim_{L \to \infty} \tilde{J}_- = \hat{J}_-.$$

Theorem B.1 implies that under weak smoothness assumptions, a sufficiently dense grid provides an arbitrarily accurate approximation to the continuously constrained optimization problem (B.1)-(B.2).

Proof of Theorem B.1. We focus on the maximization problem since the minimization problem can be analyzed analogously.

Let (ψ_L, μ_L) denote the optimal solution to the maximization version of (B.3)-(B.4). De-

fine $g(\psi, \mu; \nu) = [g_1(\psi, \mu; \nu), g_2(\psi, \mu; \nu), -g_2(\psi, \mu; \nu)]$, so that $g(\psi, \mu; \nu) \leq 0$ componentwise.

Define $\ell(\nu) := \arg\min_\ell |\nu - \nu_\ell|$. Then

$$\sup_{\nu \in [0,1]} |g(\psi, \mu; \nu) - g(\psi, \mu; \ell(\nu))| \leq C/(L-1)$$

componentwise and $g(\psi, \mu; \nu_\ell) \leq 0$ implies that

$$g(\psi, \mu; \nu) \leq C/(L-1)$$

componentwise uniformly over $\nu \in [0, 1]$. Therefore, (ψ_L, μ_L) is a feasible solution to

(B.6)
$$J_+^* := \max_{\psi, \mu} f(\psi)$$

subject to the new constraint:

$$g(\psi, \mu; \nu) \leq C/(L-1) \text{ for all } \nu \in [0, 1], \ \nu = \text{rational, and } \ n^{1/2}(\bar{X} - \mu) \in \mathcal{S}.$$

Consequently, $J_+^* \geq \tilde{J}_+ \geq J_+$, where $J_+ = \max f(\psi)$ subject to $g_1(\psi, \mu, x) \leq 0$, $g_2(\psi, \mu, x) = 0$, and $\psi \in \Psi$. Define

$$\Pi := \Big\{ \xi \geq 1, \eta : \text{there is } (\psi, \mu) \text{ such that } n^{1/2}(\bar{X} - \mu) \in \mathcal{S}, \ f(\psi) \leq \eta,$$
$$\text{and } g(\varphi, \mu; \nu) \leq C/\xi \text{ for all } \nu \in [0, 1] \Big\}.$$

Note that Π is a closed set. Therefore, by Proposition 3.3 of Jeyakumar and Wolkowicz (1990), $J_+^* \to J_+$ as $L \to \infty$ if the constraints are restricted to rational values of $\nu \in [0, 1]$. It follows from continuity of g as a function of ν that the constraints hold for all $\nu \in [0, 1]$. $\quad\square$

C Minsker's (2015) Median of Means Method

In this appendix, we describe how to carry out inference based on Minsker (2015). In particular, we consider two versions of the median of means: the one based on geometric median and the other using coordinate-wise medians. Lugosi and Mendelson (2019) propose a different version of the median of means estimator that has theoretically better properties but is more difficult to compute.

First, for the case of geometric median, let $\alpha_* := 7/18$ and $p_* := 0.1$. Define

$$\psi(\alpha_*; p_*) = (1 - \alpha_*) \log \frac{1 - \alpha_*}{1 - p_*} + \alpha_* \log \frac{\alpha_*}{p_*}.$$

Let $0 < \delta < 1$ be the level of the confidence set and set

(C.1)
$$k := \left\lfloor \frac{\log(1/\delta)}{\psi(\alpha_*; p_*)} \right\rfloor + 1.$$

Assume that δ is small enough that $k \leq n/2$. Divide the sample $X_1, \ldots, X_n$ into k disjoint groups $G_1, \ldots, G_k$ of size $\left\lfloor \frac{n}{k} \right\rfloor$ each, and define

$$\hat{\mu}_j := \frac{1}{|G_j|} \sum_{i \in G_j} X_i, \; j = 1, \ldots, k,$$

$$\hat{\mu} := \text{G.med}(\hat{\mu}_1, \ldots, \hat{\mu}_k),$$

where G.med refers to the geometric median. See Minsker (2015) and references therein for details on the geometric median. The intuition behind $\hat{\mu}$ is that it is a robust measure of the population mean vector μ since each subsample mean vector $\hat{\mu}_j$ is an unbiased estimator for μ and the aggregation method via the geometric median is robust to outliners. Because of this feature, it turns out that the finite sample bound for the Euclidean norm distance between $\hat{\mu}$ and μ depends only on $\text{tr}(\Sigma)$, but not on the higher moments (see Corollary 4.1 of

Minsker, 2015). This is the main selling point of the median of means since the finite sample probability bound for the usual sample mean assumes the existence of a higher moment (e.g. the third absolute moment in Bentkus (2003) and Theorem 2.2 in Section 2.3).

Second, Minsker (2015) also considered using coordinate-wise medians instead of using the geometric median. In this case, let $\alpha_* = 1/2$ and $p_* = 0.12$. Then k is redefined via (C.1). Let $\hat{\mu}_*$ denote the vector of coordinate-wise medians.

To estimate $\operatorname{tr}(\Sigma)$, Minsker (2015) proposed the following:

$$\hat{T}_j := \frac{1}{|G_j|} \sum_{i \in G_j} \|X_i - \hat{\mu}_j\|^2, \ j = 1, \dots, k,$$

$$\hat{T} := \operatorname{med}(\hat{T}_1, \dots, \hat{T}_k).$$

where $\|a\|$ is the Euclidean norm of a vector a. Let $B(h, r)$ denote the ball of radius r centered at h and let

$$r_n := 11\sqrt{2}\sqrt{\hat{T}\frac{\log(1.4/\delta)}{n}},$$

$$r_{n,*} := 4.4\sqrt{2}\sqrt{\hat{T}\frac{\log(1.6d_\mu/\delta)}{n - 2.4\log(1.6d_\mu/\delta)}},$$

where d_μ is the dimension of μ.

Lemma C.1 (Minsker (2015)). *Assume that*

(C.2) $$15.2\sqrt{\frac{\mathbb{E}\|X - \mu\|^4 - (tr(\Sigma))^2}{(tr(\Sigma))^2}} \leq \left(\frac{1}{2} - 178\frac{\log(1.4/\delta)}{n}\right)\sqrt{\frac{n}{\log(1.4/\delta)}}.$$

Then

(C.3) $$\mathbb{P}\left[\mu \in B(\hat{\mu}, r_n)\right] \geq 1 - 2\delta \quad and \quad \mathbb{P}\left[\mu \in B(\hat{\mu}_*, r_{n,*})\right] \geq 1 - 2\delta.$$

Proof of Lemma C.1. The result on the geometric median is the exactly the same as Corollary 4.2 of Minsker (2015). The case for the vector of coordinate-wise medians follows from combining equation (4.4) in Minsker (2015) with Proposition 4.1 of Minsker (2015). □

Lemma C.1 indicates that $\mathcal{S}$ in our setup can be chosen as

$$(\hat{\mu} - \mu)'(\hat{\mu} - \mu) \leq r_n^2,$$

or

$$(\hat{\mu}_* - \mu)'(\hat{\mu}_* - \mu) \leq r_{n,*}^2,$$

either of which gives the bound with probability at least $1 - 2\delta$. The former produces a tighter bound than the latter only when the dimension of μ is sufficiently high. Note that (C.2) requires the existence of fourth moments due to the fact that $\text{tr}(\Sigma)$ is estimated by the median of means as well. The inequality in (C.2) is a relatively mild condition when n is large.

D   An Additional Empirical Example

D.1   Bounding the Average of Log Weekly Wages

Let Y_i^* denote the log weekly wage and Z_i the years of schooling. Suppose that

(D.1) $$Y_i^* = h(Z_i) + e_i,$$

where $h : \mathbb{R} \mapsto \mathbb{R}$ is an unknown function and the error term e_i satisfies $E[e_i|Z_i] = 0$ almost surely.

In the 1960 and 1970 U.S. censuses, hours of work during a week are measured in brackets.

As a result, weekly wages are only available in terms of interval data. In view of this, assume that for each individual i, we do not observe Y_i^*, but only the interval data $[L_i, U_i]$ such that $Y_i^* \in [L_i, U_i]$ along with Z_i. Here, L_i and U_i are random variables.

Assume that the support of Z_i is finite, that is, $\mathcal{Z} \in \{z_1, \ldots, z_J\}$. Denote the values of $h(\cdot)$ on $\mathcal{Z}$ by $\{\psi_1, \ldots, \psi_J\}$. That is, $h(u) = \sum_{j=1}^{J} \psi_j 1(u = z_j)$ for $u \in \mathcal{Z}$. Suppose that the object of interest is the value of $\psi^* \equiv h(z^*)$, where z^* is not in the support of Z_i but $z_{j-1} < z^* < z_j$ for some j. This type of extrapolation problem is given as a motivating example in Manski (2007a, pp. 4-5).

To partially identify ψ^*, assume that $h(\cdot)$ is monotone non-decreasing. Specifically, we impose the monotonicity on $\mathcal{Z} \cup \{z^*\}$. That is, $h(z_1) \leq h(z_2)$ whenever $z_1 \leq z_2$ for any $z_1, z_2 \in \mathcal{Z} \cup \{z^*\}$. In addition, we have the following inequality constraints:

(D.2) $$E[L_i | Z_i = z_j] \leq \psi_j \leq E[U_i | Z_i = z_j]$$

for any $1 \leq j \leq J$. Note that (D.2) alone does not provide a bounded interval on ψ^* since z^* is not in $\mathcal{Z}$. The monotonicity assumption combined with (D.2) provides an informative bound on ψ^*.

To write the optimization problem in our canonical form, let μ denote the population

moments of the following $\bar{X}$:

$$n\bar{X} = \begin{pmatrix} \sum_{i=1}^{n} L_i 1(Z_i = z_1) \\ \vdots \\ \sum_{i=1}^{n} L_i 1(Z_i = z_J) \\ \sum_{i=1}^{n} U_i 1(Z_i = z_1) \\ \vdots \\ \sum_{i=1}^{n} U_i 1(Z_i = z_J) \\ \sum_{i=1}^{n} 1(Z_i = z_1) \\ \vdots \\ \sum_{i=1}^{n} 1(Z_i = z_J) \end{pmatrix}.$$

Then, we can rewrite the constraints (D.2) in a bilinear form:

(D.3) $$E[L_i 1(Z_i = z_j)] \leq \psi_j E[1(Z_i = z_j)] \leq E[U_i 1(Z_i = z_j)]$$

for any $1 \leq j \leq J$.

D.2 Empirical Results

We use a sample extract from the U.S. 2000 census to estimate ψ^*. The sample is restricted to males who were 40 years old with educational attainment at least grade 10, positive wages and positive hours of work. In the 2000 census, variable WKSWORK1 is the hours of work in integer value, whereas WKSWORK2 is weeks worked in intervals. As a parameter of interest, we focus on the average log wage for one year of college. To illustrate our method, in our estimation procedure, we drop all observations with those with one year of college. The resulting sample size is $n = 15,647$.

In estimation, the latent Y_i^* is log(INCWAGE/WKSWORK1), where INCWAGE is total pre-

Table 5: 95% Confidence Intervals

$\mathcal{S}$	Bounds
Box	$[6.01, 7.86]$
Chi-Square	$[6.33, 6.90]$
Sub-Gaussian-1	$[6.25, 7.10]$
Sub-Gaussian-2	$[2.62, 12.69]$
Sub-Gaussian-3	$[4.24, 11.77]$
Vector-Sub-Gaussian	$[6.32, 6.87]$
Minsker	$[2.60, 12.69]$
Oracle	$[6.56, 6.61]$

tax wage and salary income and `WKSWORK1` is weeks worked in integer value. We observe brackets $[L_i, U_i]$ of log(`INCWAGE`/`WKSWORK2`), where `WKSWORK2` is weeks worked in intervals: [1,13], [14,26], [27,39], [40,47], [48,49], [50,52]. The brackets $[L_i, U_i]$ are random because the numerator `INCWAGE` is random. The set $\mathcal{Z}$ includes 7 different values: 10th grade, 11th grade, 12th grade, 1 year of college, 2 years of college, 4 years of college and 5 or more years of college. Since we drop observations with those with 1 year of college, we have a missing data problem as described in the previous section.

Table 5 shows estimation results. The nominal level of coverage is 0.95. Inference methods are the same as those used in Section 4. The `Oracle` refers to the confidence interval of the average log wage for one year of college using the actual value of weekly wages for those with one year of college. As in the empirical example in the main text, the `Chi-Square` and `Vector-Sub-Gaussian` intervals are tighter than the other intervals and are only slightly wider than the `Oracle` interval.

References

ANGRIST, J. D., AND W. N. EVANS (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 88(3), 450–477.

BELLONI, A., F. BUGNI, AND V. CHERNOZHUKOV (2018): "Subvector Inference in Partially

Identified Models with Many Moment Inequalities," arXiv:1806.11466 [math.ST] `https://arxiv.org/abs/1806.11466`.

BENTKUS, V. (2003): "On the dependence of the Berry–Esseen bound on dimension," *Journal of Statistical Planning and Inference*, 113(2), 385–402.

BERTSIMAS, D., A. KING, AND R. MAZUMDER (2016): "Best subset selection via a modern optimization lens," *Annals of Statistics*, 44(2), 813–852.

BLUNDELL, R., A. DUNCAN, AND C. MEGHIR (1998): "Estimating Labor Supply Responses Using Tax Reforms," *Econometrica*, 66(4), 827–861.

BUGNI, F. A., I. A. CANAY, AND X. SHI (2017): "Inference for subvectors and other functions of partially identified parameters in moment inequality models," *Quantitative Economics*, 8(1), 1–38.

BÜHLMANN, P., AND S. VAN DE GEER (2011): *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media.

CANAY, I. A., AND A. M. SHAIKH (2017): "Practical and Theoretical Advances in Inference for Partially Identified Models," in *Advances in Economics and Econometrics: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, vol. 2 of *Econometric Society Monographs*, pp. 271–306. Cambridge University Press.

CHEN, X., T. M. CHRISTENSEN, AND E. TAMER (2018): "Monte Carlo Confidence Sets for Identified Sets," *Econometrica*, 86(6), 1965–2018.

CHERNOZHUKOV, V., D. CHETVERIKOV, AND K. KATO (2017): "Central limit theorems and bootstrap in high dimensions," *Annals of Probability*, 45(4), 2309–2352.

——— (2019): "Inference on causal and structural parameters using many moment inequalities," *Review of Economic Studies*, 86(5), 1867–1900.

CHERNOZHUKOV, V., C. HANSEN, AND M. JANSSON (2009): "Finite sample inference for quantile regression models," *Journal of Econometrics*, 152(2), 93–103.

CILIBERTO, F., AND E. TAMER (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77(6), 1791–1828.

CLOPPER, C. J., AND E. S. PEARSON (1934): "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, 26(4), 404–413.

DASGUPTA, A. (2008): *Asymptotic Theory of Statistics and Probability.* Springer, New York.

FREYBERGER, J., AND J. L. HOROWITZ (2015): "Identification and shape restrictions in nonparametric instrumental variables estimation," *Journal of Econometrics*, 189(1), 41–53.

Ho, K., and A. Pakes (2014): "Hospital Choices, Hospital Prices, and Financial Incentives to Physicians," *American Economic Review*, 104(12), 3841–84.

Ho, K., and A. M. Rosen (2017): "Partial Identification in Applied Research: Benefits and Challenges," in *Advances in Economics and Econometrics: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, vol. 2 of *Econometric Society Monographs*, pp. 307–359. Cambridge University Press.

Horowitz, J. L., and S. Lee (2017): "Nonparametric estimation and inference under shape restrictions," *Journal of Econometrics*, 201(1), 108–126.

Hsieh, Y.-W., X. Shi, and M. Shum (2017): "Inference on Estimators defined by Mathematical Programming," arXiv:1709.09115 [econ.EM], https://arxiv.org/abs/1709.09115.

Hsu, D., S. Kakade, and T. Zhang (2012): "A tail inequality for quadratic forms of subgaussian random vectors," *Electronic Communications in Probability*, 17, 1 – 6.

Jeyakumar, V., and H. Wolkowicz (1990): "Zero duality gaps in infinite-dimensional programming," *Journal of Optimization Theory and Applications*, 67(1), 87–108.

Kaido, H., F. Molinari, and J. Stoye (2019): "Confidence Intervals for Projections of Partially Identified Parameters," *Econometrica*, 87(4), 1397–1432.

Kline, B., and E. Tamer (2016): "Bayesian inference in a class of partially identified models," *Quantitative Economics*, 7(2), 329–366.

Kline, P., and M. Tartari (2016): "Bounding the Labor Supply Responses to a Randomized Welfare Experiment: A Revealed Preference Approach," *American Economic Review*, 106(4), 972–1014.

Lugosi, G., and S. Mendelson (2019): "Sub-Gaussian estimators of the mean of a random vector," *Annals of Statistics*, 47(2), 783–794.

Manski, C. F. (2007a): *Identification for Prediction and Decision*. Harvard University Press, Cambridge, Massachusetts.

———— (2007b): "Partial Identification of Counterfactual Choice Probabilities," *International Economic Review*, 48(4), 1393–1410.

Minsker, S. (2015): "Geometric median and robust estimation in Banach spaces," *Bernoulli*, 21(4), 2308–2335.

Molinari, F. (2020): "Microeconometrics with Partial Identification," arXiv:2004.11751 [econ.EM] https://arxiv.org/abs/2004.11751.

Raič, M. (2019): "A multivariate Berry-Esseen theorem with explicit constants," *Bernoulli*, 25(4A), 2824–2853.

REGUANT, M. (2016): "Bounding Outcomes in Counterfactual Analysis," Northwestern University Working Paper.

RIGOLLET, P., AND J.-C. HÜTTER (2017): "High Dimensional Statistics," Lecture Notes available at http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf.

ROSEN, A. M., AND T. URA (2019): "Finite Sample Inference for the Maximum Score Estimand," arXiv:1903.01511 [econ.EM] https://arxiv.org/abs/1903.01511.

RUGGLES, S., S. FLOOD, S. FOSTER, R. GOEKEN, J. PACAS, M. SCHOUWEILER, AND M. SOBEK (2021): "IPUMS USA: Version 11.0 [dataset]," Minneapolis, MN https://doi.org/10.18128/D010.V11.0.

SHI, X., AND M. SHUM (2015): "Simple Two-stage Inference for a Class of Partially Identified Models," *Econometric Theory*, 31(3), 493–520.

SPOKOINY, V., AND M. ZHILOVA (2015): "Bootstrap confidence sets under model misspecification," *Annals of Statistics*, 43(6), 2653–2675.

SYRGKANIS, V., E. TAMER, AND J. ZIANI (2018): "Inference on Auctions with Weak Assumptions on Information," arXiv:1710.03830 [econ.EM], https://arxiv.org/abs/1710.03830.

TAMER, E. (2010): "Partial Identification in Econometrics," *Annual Review of Economics*, 2(1), 167–195.

VERSHYNIN, R. (2018): *High-dimensional probability: An introduction with applications in data science.* Cambridge university press.

WAINWRIGHT, M. J. (2019): *High-dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge University Press.

ZHILOVA, M. (2020): "Nonclassical Berry-Esseen inequalities and accuracy of the bootstrap," *Annals of Statistics*, 48(4), 1922–1939.